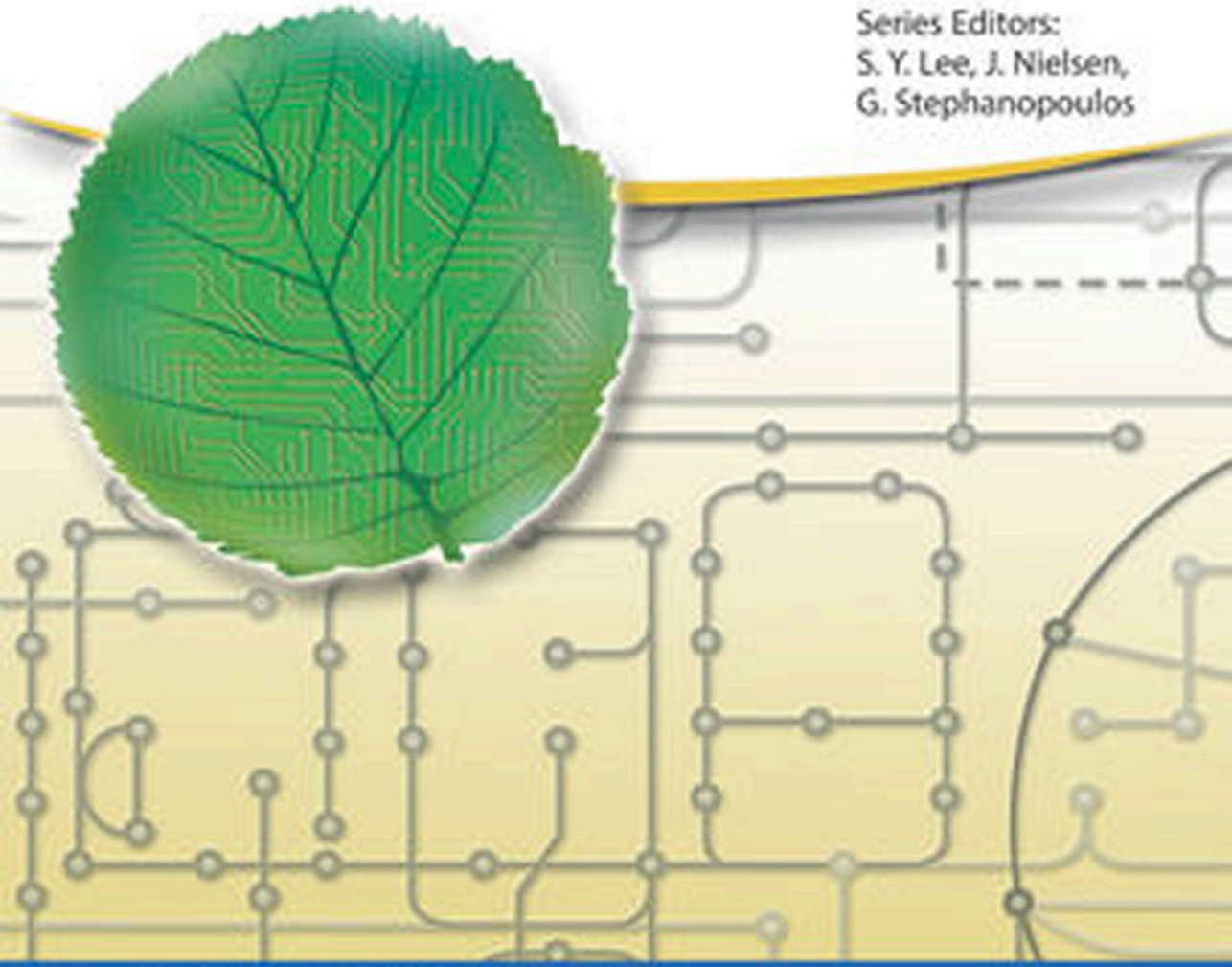
WILEY-VCH

Edited by Christina Smolke

Synthetic Biology

Parts, Devices and Applications

Volume 8
Series Editors:
S. Y. Lee, J. Nielsen,
G. Stephanopoulos



Advanced Biotechnology

Synthetic Biology: Parts, Devices and Applications

Related Titles

Luisi, P.P. (ed.)

Chemical Synthetic Biology

2011
Print ISBN: 978-0-470-71397-6

Further Volumes of the "Advanved Biotechnology" Series:

Published:

Villadsen, J. (ed.)

Fundamental Bioengineering

2016
Print ISBN: 978-3-527-33674-6

Love, J. Ch. (ed.)

Micro- and Nanosystems for Biotechnology

2016
Print ISBN: 978-3-527-33281-6

Wittmann, Ch., Liao, J.C. (eds.)

Industrial Biotechnology

Microorganisms (2 Volumes)

2017
Print ISBN: 978-3-527-34179-5

Wittmann, Ch., Liao, J.C. (eds.)

Industrial Biotechnology

Products and Processes

2017
Print ISBN: 978-3-527-34181-8

Yoshida, T. (ed.)

Applied Bioengineering

2017
Print ISBN: 978-3-527-34075-0

Nielsen, J., Hohmann, S. (eds)

Systems Biology

2017
Print ISBN: 978-3-527-33558-9

Coming soon:

Chang, H. N.

Emerging Areas in Bioengineering

2018
Print ISBN: 978-3-527-34088-0

Planned:

Lee, G. M., Faustrup Kildegaard, H.

Cell Culture Engineering

Recombinant Protein Production

2018
Print ISBN: 9783527343348

Synthetic Biology: Parts, Devices and Applications

Edited by Christina Smolke

WILEY-VCH

Volume Editor

Christina Smolke
Stanford School of Medicine
Y2E2 Building, Room #B07
473 Via Ortega
Stanford, CA 94305
United States

Series Editors

Sang Yup Lee
KAIST
Department of Chemical & Biomolecular Engineering
291 Daehak-ro, Yuseong-gu
34141 Daejeon
Republic of Korea

Jens Nielsen
Chalmers University of Technology
Department of Biology and Biological Engineering
Kemivägen 10
41296 Göteborg
Sweden

Gregory Stephanopoulos
Massachusetts Institute of Technology
Department of Chemical Engineering
77 Massachusetts Avenue
Cambridge, MA 02139
USA

Cover
fotolia_yurakp and fotolia_EvgeniyBobrov

Library of Congress Card No.: applied for

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at <http://dnb.d-nb.de.>

Print ISBN: 978-3-527-33075-1
ePDF ISBN: 978-3-527-68808-1
ePub ISBN: 978-3-527-68809-8
Mobi ISBN: 978-3-527-68807-4
oBook ISBN: 978-3-527-68810-4

Cover Design Adam-Design, Weinheim, Germany
Typesetting SPi Global, Chennai, India
Printing and Binding

Printed on acid-free paper

Contents

About the Series Editors

Sang Yup Lee is distinguished Professor at the Department of Chemical and Biomolecular Engineering at the Korea Advanced Institute of Science and Technology. At present, Prof. Lee is the Director of the Center for Systems and Synthetic Biotechnology, Director of the BioProcess Engineering Research Center, and Director of the Bioinformatics Research Center. He has published more than 500 journal papers, 64 books, and book chapters, and has more than 580 patents (either registered or applied) to his credit. He has received numerous awards, including the National Order of Merit, the Merck Metabolic Engineering Award, the ACS Marvin Johnson Award, Charles Thom Award, Amgen Biochemical Engineering Award, Elmer Gaden Award, POSCO TJ Park Prize, and HoAm Prize. He is Fellow of American Association for the Advancement of Science, the American Academy of Microbiology, American Institute of Chemical Engineers, Society for Industrial Microbiology and Biotechnology, American Institute of Medical and Biological Engineering, the World Academy of Science, the Korean Academy of Science and Technology, and the National Academy of Engineering of Korea. He is also Foreign Member of National Academy of Engineering, USA. In addition, he is honorary professor of the University of Queensland (Australia), honorary professor of the Chinese Academy of Sciences, honorary professor of Wuhan University (China), honorary professor of Hubei University of Technology (China), honorary professor of Beijing University of Chemical Technology (China), and advisory professor of the Shanghai Jiaotong University (China). Apart from his academic associations, Prof. Lee is the editor-in-chief of the *Biotechnology Journal* and is also contributing to numerous other journals as associate editor and board member. Prof. Lee is serving as a member of Presidential Advisory Committee on Science and Technology (South Korea).

Jens Nielsen is Professor and Director to Chalmers University of Technology (Sweden) since 2008. He obtained an MSc degree in chemical engineering and a PhD degree (1989) in biochemical engineering from the Technical University of Denmark (DTU) and after that established his independent research group and was appointed full professor there in 1998. He was Fulbright visiting professor at MIT in 1995–1996. At DTU, he founded and directed the Center for Microbial Biotechnology. Prof. Nielsen has published more than 350 research papers and coauthored more than 40 books, and he is inventor of more than 50 patents. He has founded several companies that have raised more than 20 million in venture capital. He has received numerous Danish and international awards and is member of the Academy of Technical Sciences (Denmark), the National Academy of Engineering (USA), the Royal Danish Academy of Science and Letters, the American Institute for Medical and Biological Engineering and the Royal Swedish Academy of Engineering Sciences.

Gregory Stephanopoulos is the W.H. Dow Professor of Chemical Engineering at the Massachusetts Institute of Technology (MIT, USA) and Director of the MIT Metabolic Engineering Laboratory. He is also Instructor of Bioengineering at Harvard Medical School (since 1997). He received his BS degree from the National Technical University of Athens and PhD from the University of Minnesota (USA). He has coauthored about 400 research papers and 50 patents, along with the first textbook on metabolic engineering. He has been recognized by numerous awards from the American Institute of Chemical Engineers (AIChE) (Wilhelm, Walker and Founders awards), American Chemical Society (ACS), Society of Industrial Microbiology (SIM), BIO (Washington Carver Award), the John Fritz Medal of the American Association of Engineering Societies, and others. In 2003, he was elected member of the National Academy of Engineering (USA) and in 2014 President of AIChE.

Part I

DNA Synthesis and Genome Engineering

1

Competition and the Future of Reading and Writing DNA

Robert Carlson

Biodesic and Bioeconomy Capital, 3417 Evanston Ave N, Ste 329, Seattle, WA 98103, USA

Constructing arbitrary genetic instruction sets is a core technology for biological engineering. Biologists and engineers are pursuing even better methods to assemble these arbitrary sequences from synthetic oligonucleotides (oligos) [1]. These new assembly methods in principle reduce costs, improve access, and result in long sequences of error-free DNA that can be used to construct entire microbial genomes [2]. However, an increasing diversity of assembly methods is not matched by any obvious corresponding innovation in producing oligos. Commercial oligo production employs a very narrow technology base that is many decades old. Consequently, there is only minimal price and product differentiation among corporations that produce oligos. Prices have stagnated, which in turn limits the economic potential of new assembly methods that rely on oligos. Improvements may come via recently demonstrated assembly methods that are capable of using oligos of lower quality and lower cost as feedstocks. However, while these new methods may substantially lower the cost of gene-length double-stranded DNA (dsDNA), they also may be economically viable only when producing many orders of magnitude with more dsDNA than what is now used by the market. The commercial success of these methods, and the broader access to dsDNA they enable, may therefore depend on structural changes in the market that are yet to emerge.

1.1 Productivity Improvements in Biological Technologies

In considering the larger impact of technological monoculture in DNA synthesis, it is useful to contrast DNA synthesis and assembly with DNA sequencing. In particular, it is instructive to compare productivity estimates of commercially available sequencing and synthesis instruments (Figure 1.1). Reading DNA is as crucial as writing DNA to the future of biological engineering. Due to not just commercial competition but also competition between sequencing technologies, both

Synthetic Biology: Parts, Devices and Applications, First Edition. Edited by Christina Smolke.

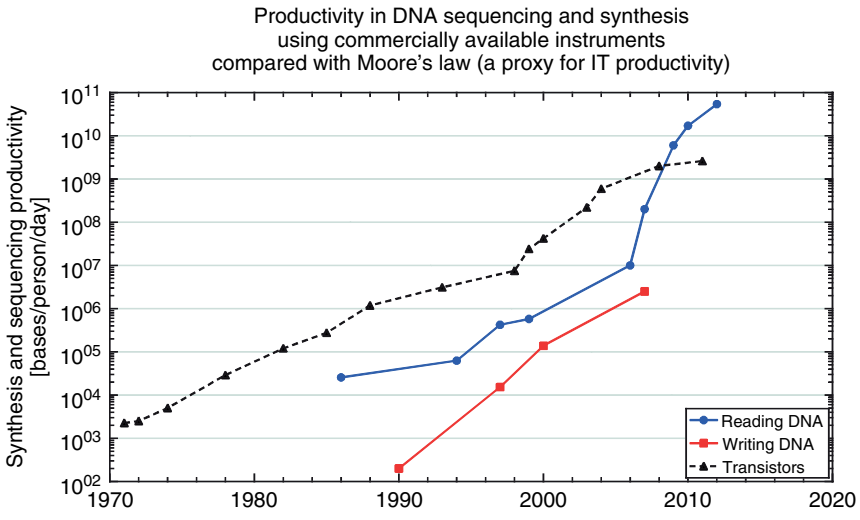


Figure 1.1 Estimates of the maximum productivity of DNA synthesis and sequencing enabled by commercially available instruments. Productivity of DNA synthesis is shown only for column-based synthesis instruments, as data for sDNA fabricated on commercially available DNA arrays is unavailable; exceptions are discussed in the text. Shown for comparison is Moore's law, the number of transistors per chip. (Intel; Carlson, 2010 [3]; Loman *et al*. 2012 [4]; Quail *et al*. 2012 [5]; Liu, 2012 [6].)

prices and instrument capabilities are improving rapidly. The technological diversity responsible for these improvements poses challenges in making quantitative comparisons. As in previous discussions of these trends, in what follows I rely on the metrics of price [$/base] and productivity [bases/person/day].

Figure 1.1 also directly compares the productivity enabled by commercially available sequencing and synthesis instruments to Moore's law, which describes the exponential increase in transistor counts in CPUs over time. Readers new to this discussion are referred to References 3 and 4 for in-depth descriptions of the development of these metrics and the utility of a comparison with Moore's law [3, 7]. Very briefly, Moore's law is a proxy for productivity; more transistors enable greater computational capability, which putatively equates to greater productivity.

Visual inspection of Figure 1.1 reveals several interesting features. First, general synthesis productivity has not improved for several years because no new instruments have been released publicly since about 2008. Productivity estimates for instruments developed and run by oligo and gene synthesis service providers are not publicly available.[1]

1 It is likely that array-based DNA synthesis used to supply gene assembly operates at a much higher productivity than column-based synthesis. For example, Agilent reportedly produces and ships in excess of 30 billion bases of ssDNA a day, the equivalent of more than 10 human genomes, on an undisclosed number of arrays (Darlene Solomon, Personal Communication).

Second, it is clear that DNA sequencing platforms are improving very rapidly, now much faster than Moore's law.

Moore's law and its economic and social consequences are often used to benchmark our expectations of other technologies. Therefore, developing an understanding of this "law" provides a means to compare and contrast it with other technological trends.

1.2 The Origin of Moore's Law and Its Implications for Biological Technologies

Moore's law is often mistakenly described as a technological inevitability or is assumed to be some sort of physical phenomenon. It is neither; Moore's law is a business plan, and as such it is based on economics and planning. Gordon Moore's somewhat opaque original statement of what became the "law" was a prediction concerning economically viable transistor yields [8]. Over time, Moore's economic observation became an operational model based on monopoly pricing, and it eventually enabled Intel to outcompete all other manufacturers of general CPUs. Two important features distinguish CPUs from other technologies and provide insight into the future of trends in biological technologies: the first is the cost of production, and the second is the monopoly pricing structure.

Early on Intel recognized the utility of exploiting Moore's law as a business plan. A simple scaling argument reveals the details of the plan. While transistor counts increased exponentially, Intel correspondingly reduced the price per transistor at a similar rate. In order to maintain revenues, the company needed to ship proportionally more transistors every quarter; in fact, the company increased its shipping numbers faster than prices fell, enabling consistent revenue to grow for several decades. This explains why Intel former CEO Andy Grove reportedly constantly pushed for an even greater scale [9].

In this sense, Moore's law was always about economics and planning in a multibillion-dollar industry. In the year 2000, a new chip fab cost about $1 billion; in 2009, it cost about $3 billion. Now, according to *The Economist*, Intel estimates that a new chip fab costs about $10 billion [9]. This apparent exponential increase in the cost of semiconductor processing is known as Rock's law. It is often argued that Moore's law will eventually expire due to the physical constraints of fabricating transistors at small length scales, but it is more likely to become difficult to economically justify constructing fabrication facilities at the cost of tens to hundreds of billions of dollars. Even through the next several iterations, these construction costs will dictate careful planning that spans many years. No business spends $10 billion without a great deal of planning, and, more directly, no business finances a manufacturing plant that expensive without demonstrating a long-term plan to repay the financiers. Moreover, Intel must coordinate the manufacturing and delivery of very expensive, very complex semiconductor processing instruments made by other companies. Thus Intel's planning and finance cycles explicitly extend many years into the future. New technology has certainly been required to achieve each planning goal, but this is part of the ongoing research, development, and planning process for Intel.

Moore's law served a second purpose for Intel and one that is less well recognized but arguably more important; it was a pace selected to enable Intel to win. Intel successfully organized an entire industry to move at a pace only it could survive. And only Intel did survive. While Intel still has competitors in products such as memory or GPUs, companies that produced high volume, general CPUs have all succumbed to the pace of Moore's law. The final component of this argument is that, according to Gordon Moore, Intel could have increased transistor counts faster than the historical rate.[2] In fact, Intel ran on a faster internal innovation clock than it admitted publicly, which means that Moore's law was, as one Intel executive put it, a "marketing head fake" [10]. The inescapable conclusion of this argument is that the management of Intel made a very careful calculation; they evaluated product rollouts to consumers – the rate of new product adoption, the rate of semiconductor processing improvements, and the financial requirements for building the next chip fab line – and then set a pace that nobody else could match but that left Intel plenty of headroom for future products. In effect, if not intent, Intel executed a strategy that enabled it to set CPU prices and then to reduce those prices at a rate no other company could match.

This long-term planning, pricing structure, and the resulting lack of competition contrasts quite strongly with the commercial landscape for biological technologies. Whereas the exponential pace of doubling of transistor counts was controlled by just one company, productivity in DNA sequencing has recently improved faster than Moore's law due to competition not just among companies but also among technologies. Conversely, the lack of improvement in synthesis productivity suggests that the narrow technology base for writing DNA has reached technical and, therefore, economic limits. Nonetheless, while Figure 1.1 may suggest a temporary slowdown in the rate of improvement for sequencing, and in effect shows zero recent improvement for synthesis, new technologies will inevitably facilitate continued competition and, therefore, continued productivity improvement.

1.3 Lessons from Other Technologies

Compared with that in other industries, the financial barrier to entry in biological technologies is quite low. Unlike chip manufacturing, there is nothing in biology with a commercial development price tag of $10 billion. The Boeing 787 reportedly cost $32 billion to develop as of 2011 and is on top of a century of multibillion-dollar aviation projects that preceded it [11]. Better Place, an electric car company, declared bankruptcy after receiving $850 million in investment [12]. Tesla Motors has reported only one profitable quarter since 2003 and continues to operate in the red while working to achieve manufacturing scale-up [13, 14].

There are two kinds of costs that are important to distinguish here. The first is the cost of developing and commercializing a particular product. Based on the

2 Gordon Moore to Danny Hillis, as related by Danny Hillis, Personal Communication.

money reportedly raised and spent by Illumina, Pacific Biosciences, Oxford Nanopore, Life, Ion Torrent, and Complete Genomics (the latter three before acquisition), it appears that developing and marketing a second-generation sequencing technology can cost more than $100 million. Substantially more money gets spent, and lost, in operations before any of these product lines is revenue positive. Nonetheless, relatively low development costs have enabled a number of companies to enter the market for DNA sequencing, resulting in a healthy competition in a market that is presently modest in size but that is expected to grow rapidly over the coming decades.

1.4 Pricing Improvements in Biological Technologies

The second kind of cost to keep in mind is the use of new technologies to produce an object or produce data. Figure 1.2 is a plot of commercial prices for column-synthesized oligos, gene-length synthetic DNA (sDNA), and DNA sequencing. Prior to 2006, the sequencing market was dominated by Sanger-based capillary instruments produced by Applied Biosystems, in effect another pricing monopoly. After 2006, the market saw a rapid proliferation of not just commercial but also technological competition with the launch of next-generation systems from 454, Illumina, Ion Torrent, Pacific Biosciences, and Oxford Nanopore based on a diversity of chemical and physical detection methodologies [15]. Illumina presently dominates the market for sequencing instruments but is facing competition from Oxford Nanopore and various Chinese insurgents. There also remains technological diversity between companies, which contributes to competitive

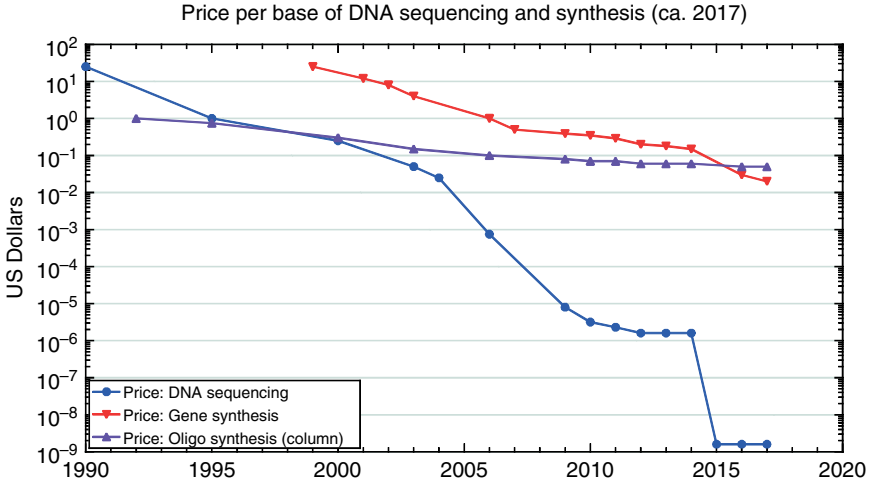


Figure 1.2 Commercial prices per base for DNA sequencing, column-synthesized oligonucleotides, and gene-length sDNA. Reported prices for array-synthesized oligos vary widely, and no time series is available. Market pricing for genes can vary by up to an order of magnitude, depending on sequencing composition and complexity. (Carlson (2010), Commercial price quotes.)

pressures. An important consequence of the emergence of technological competition in the DNA sequencing market is a rapid price decrease. The NIH maintains a version of this plot that compares sequencing prices with cost per megabyte for memory, another form of Moore's law [16]. Both Figure 1.2 and the NIH plot show that sequencing costs kept pace with Moore's law while a pricing monopoly was in effect. The emergence of technological competition produced both productivity improvements and price changes that outpaced Moore's law.

In contrast, despite modest commercial competition in the DNA synthesis market, the lack of technological competition has limited price decreases in the last 5 years. The industry as it exists today is based on chemistry that is several decades old, in which oligos are synthesized step by step on an immobilized substrate. Using array-synthesized oligos for gene assembly appears to be lowering the market price, though quality and delivery time are reportedly inconsistent across the industry. Improved error correction and removal technologies may further reduce the assembly cost for genes and thereby improve the profit margins [17]. My informal conversations with industry insiders suggest that oligo producers may no longer include the cost of goods in calculating prices; that is, oligo prices are evidently determined largely by the cost of capital rather than the cost of raw materials. This suggests that very little pricing improvement can be expected for genes produced from standard oligo synthesis.

1.5 Prospects for New Assembly Technologies

Array synthesis has the advantage of a low volume production of oligos with high library diversity [18]. Gene assembly based on array synthesis has proved difficult to commercialize. At least three companies in this space, Codon Devices, Gen9, and Cambrian Genomics, have gone bankrupt or been acquired in recent years. Twist, a more recent entrant, now quotes prices in the neighborhood of $10 per base and publicly asserts it will push prices much lower in the coming years.

With prices potentially soon falling by orders of magnitude, one must ask about the subsequent impact on the market for synthetic genes. New firms entering the market are implicitly working on the hypothesis that supply-side price reductions will drive increased demand. The most obvious source of that demand would be forward design of genetic circuits based on rational models. Yet the most sophisticated synthetic genetic circuits being constructed in industrial settings are designed largely using heuristic models rather than quantitative design tools [19]. Moreover, these circuits contain only a handful of components, which stand as a substantial bottleneck for demand. Alternatively, customers may employ less up-front predictive design and instead rely on high-throughput screening of pathway variants; screening libraries of pathways has the potential to create substantial demand for synthetic genes [20].

Considering the interplay between market size and price reveals challenges for companies entering the gene synthesis industry. Recalling the lessons of Moore's law, a relatively simple scaling argument will reveal the performance constraints

of the gene synthesis industry. Intel knew that it could grow financially in the context of exponentially falling transistor costs by shipping exponentially more transistors every quarter – that is, the business model of Moore's law. But that was in the context of an effective pricing monopoly, and Intel's success required a market that grew exponentially for decades. The question for synthetic gene companies is whether the market will grow fast enough to provide adequate revenues when prices fall. For every order of magnitude drop in the price of synthetic genes, the industry will have to ship an order of magnitude of more DNA just to maintain constant revenues. More broadly, in order for the industry to grow, synthesis companies must find a way to expand their market at a rate faster than when prices fall. Unfortunately, as best as I can tell, despite falling prices and putative increases in demand, the global gene synthesis industry generated only about $150 million in 2015 [21]. The total size of the industry appears to have been static, or even to have decreased, over the prior decade.

Ultimately, for a new wave of gene synthesis companies to be successful, they have to provide their customers with something of value. Academic customers are likely to become more plentiful as it becomes even more obvious that ordering genes is cheaper than cloning genes, even with graduate student labor costs. Gene synthesis pioneer John Mulligan used to cite NIH expenditures on cloning – approximately $3 billion annually – as a potential market size for gene synthesis [22]. This is certainly an attractive potential market. However, with the price per base potentially falling dramatically in the near term, the comparison to cloning must focus on the total number of cloned bases replaced by synthesis and at what exact price.

For commercial customers, it is less obvious that lower prices will equate to substantial increases in demand. The cost of sDNA is always going to be a small cost of developing a product, and it is not obvious that making a small cost even smaller will affect the operations of an average corporate lab. In general, research only accounts for 1–10% of the cost of the final product [23]. The vast majority of development costs are in scaling up production and in polishing the product into something customers will actually buy. For the sake of argument, assume that the total metabolic engineering development costs for a new product are in the neighborhood of $50–100 million, a reasonable estimate given the amounts that companies such as Gevo and Amyris have reportedly spent. In that context, reducing the cost of sDNA from $50 000 to $500 may be useful, but the corporate scientist-customer will be more concerned about reducing the $50 million overall costs by a factor of two, or even an order of magnitude, a decrease that would drive the cost of sDNA into the noise. Thus, in order to increase demand adequately, the production of radically cheaper sDNA must be coupled with innovations that reduce the overall the product development costs. As suggested above, forward design of complex circuits is unlikely to provide adequate innovation anytime soon. An alternative may be high-throughput screening operations that enable testing many variant pathways simultaneously. But note that this is not just another hypothesis about how the immediate future of engineering biology will change but also another generally unacknowledged hypothesis. It might turn out to be wrong, and elucidating one final difference between transistors and DNA may explain why.

The global market for transistors has grown consistently for decades, driven by an insatiable demand for more computational power and digital storage. Every new product must contain more transistors than the model it replaces. In contrast, while the demand for biological products is also growing, every new biological product is made using, in principle, just one DNA sequence. In practice, while many different DNA sequences may be constructed and tested in developing a new product, these many sequences are still winnowed down to only one sequence that defines a microbial, plant, or mammalian production strain. Nevertheless, this fundamental difference in use between transistors and DNA reveals the gene synthesis industry as the provider of engineering prototypes rather than as a large volume manufacturer of consumer goods. Consequently, while high-throughput synthetic biology companies such as Amyris, Ginkgo Bioworks, and Zymergen may place relatively large orders for sDNA, the price and volume of that sDNA will never have much impact on the final products produced by those companies.

1.6 Beyond Programming Genetic Instruction Sets

At present, the cost of purifying oligos and short dsDNA can exceed the cost of the DNA itself by as much a factor of three. The availability of lower cost, high quality dsDNA may therefore enable applications that are presently not economically viable at large scale. Beyond its utility in programming biological systems, dsDNA can be used as nanoscale structural or functional components [24]. The future of these applications is difficult to predict but could include circuitry assembled from DNA that is modified using proteins and chemistry to create conductive and semiconductive regions useful for computation [25]. It is unclear what sDNA market size these applications may support. Recent progress suggests that new demand might emerge from the use of DNA as a digital information storage medium [26]. Even a single, modestly size data center would consume many orders of magnitude of more sDNA than any prospective use of sDNA in biological contexts [27].

1.7 Future Prospects

Regardless of the particular course of companies entering the gene synthesis market, it appears that prices are likely to fall, potentially fueling an increase in demand. That demand may come in part from customers who fall outside the usual academic and corporate classifications; start-up companies, community labs, and individual, independent entrepreneurs and scientists are likely to use sDNA in new and interesting ways. The standing biosecurity strategy of the United States is to explicitly engage and encourage this innovation, including in contexts such as "garages and basements" [28]. This strategy recognizes the important role of entrepreneurs in innovation and job creation and also recognizes the difficulty of preventing access to biological technologies through regulations or restrictions.

Complementing the engagement strategy is an effort to prevent accidentally synthesizing and shipping potentially hazardous sequences. Most gene synthesis companies have voluntarily signed onto international agreements to screen orders against lists of pathogens and toxins such as the Harmonized Screening Protocol of the International Gene Synthesis Consortium (IGSC) [29].

The technical potential of new sDNA production methods may provide an opportunity to build and test far more genetic circuit designs than what is now feasible. The economic demand for biological production is enormous and is growing rapidly [30, 31]. Whether newly emerging sDNA companies survive economically depends in large part on their ability to increase total market demand sufficiently to offset falling prices. The size of that market, in turn, largely depends on whether less expensive dsDNA enables customers to reduce research and development costs and to create more products. The fundamental problem for the synthesis industry is that, however valuable sDNA is substantively to biological engineering in practice, the monetary value of that DNA is small compared with total development costs and has been falling, at times very rapidly, for decades. Falling prices limit both the maximum profit margin and the incentive to invest in new technology. Any new technology that does enter the market will inevitably drive competition, further depressing prices and margins. Going forward, productivity and prices are likely to display step changes resulting from the emergence of new technology and competition rather than display smooth long-term changes. Finally, given the relatively low barriers to entry for biological technologies and the consequent inevitable competition, it is worth asking whether centralized production is the future of the industry. As with printing documents, it may be that the economics of printing and using DNA favor distributed production, perhaps even a desktop model. There is no fundamental barrier to integrating any demonstrated synthesis and assembly technologies into a desktop gene printer. Ultimately, over the long term, a globally expanding customer base will ultimately determine how sDNA is produced and used. Regardless of how current technology specifically impacts supply and prices, that customer base is increasing, and it is likely that the trends displayed in Figures 1.1 and 1.2 will continue for many years to come.

References

1 Ellis, T. *et al.* (2011) DNA assembly for synthetic biology: from parts to pathways and beyond. *Integr. Biol.*, **3** (2), 109–118.

2 Baker, M. (2012) De novo genome assembly: what every biologist should know. *Nat. Methods*, **9**, 333–337.

3 Carlson, R. (2010) *Biology is Technology: The Promise, Peril, and New Business of Engineering Life*, Harvard University Press.

4 Loman, N. *et al.* (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.*, **30**, 434–439.

5 Quail, M. *et al* (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**, 341.

6 Liu, L. (2012) Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.*, Article ID 251364.
7 Carlson, R. (2004) The pace and proliferation of biological technologies. *Biosecur. Bioterror.*, **1** (3), 203–214.
8 Moore, G.E. (1965) Cramming more components onto integrated circuits. *Electronics*, 114–117.
9 Under new management, The Economist, 2 April 2009.
10 Hutcheson, G.D. (2009) The economic implications of Moore's Law, in *Into the Nano Era: Moore's Law Beyond Planar Silicon CMOS* (ed. H. Huff), Springer.
11 Wikipedia Boeing 787 Dreamliner, http://en.wikipedia.org/wiki/Boeing_787_Dreamliner (accessed 28 May 2013).
12 Wired Better Place Runs Out of Juice, Reportedly Plans Bankruptcy, http://www.wired.com/autopia/2013/05/better-place-bankruptcy-report/ (accessed 10 January 2018).
13 Higgins, T. (2016) Tesla Generates a Profit. The Wall Street Journal (Oct. 26).
14 Reuters (2016) Tesla Just Raised $1.5 Billion to Finance Model 3 Production (May 20).
15 See Carlson, R. (2013) How Competition Improves DNA Sequencing, http://www.synthesis.cc/synthesis/2013/04/how_competition_improves_reading_dna (accessed 3 January 2018).
16 National Institutes of Health DNA Sequencing Costs, http://www.genome.gov/sequencingcosts/ (accessed 10 January 2018).
17 Carlson, R. (2009) The changing economics of DNA synthesis. *Nat. Biotechnol.*, **27** (12), 1091–1094.
18 Tian, J. *et al.* (2004) Accurate multiplex gene synthesis from programmable DNA microchips. *Nature*, **432** (7020), 23.
19 Carlson (2009, 2010); Paddon, C.J. *et al.* (2013) High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature*, **496** (7446).
20 See, for example, Integrated DNA Technologies Building Biological Factories for Renewable and Sustainable Products, IDT Decoded Newsletter, http://www.idtdna.com/pages/decoded/decoded-articles/your-research/decoded/2013/01/18/building-biological-factories-for-renewable-and-sustainable-products (accessed 10 January 2018).
21 Carlson, R. (2016) On DNA and Transistors http://www.synthesis.cc/synthesis/2016/03/on_dna_and_transistors (accessed 3 January 2018).
22 See, for example, Carlson, R. (2009) On the Demise of Codon Devices, http://www.synthesis.cc/synthesis/2009/04/on_the_demise_of_condon_devices (accessed 3 January 2018).
23 See, for example, National Research Council (1999) *Funding a Revolution: Government Support for Computing Research*, National Academies Press, Washington, DC.
24 (a) Tørring, T. *et al.* (2011) DNA origami: a quantum leap for self-assembly of complex structures. *Chem. Soc. Rev.*, **12**, 5636; (b) Venkataraman, S. *et al.* (2007) An autonomous polymerization motor powered by DNA hybridization. *Nat. Nanotechnol.*, **2**, 490; (c) Gu, H. (2010) A proximity-based programmable DNA nanoscale assembly line. *Nature*, **465**, 202–205.

25 (a) Kershner, R. *et al.* (2009) Placement and orientation of individual DNA shapes on lithographically patterned surfaces. *Nat. Nanotechnol.*, **4**, 55; (b) Keren, K. *et al.* (2003) DNA-templated carbon nanotube field-effect transistor. *Science*, **302** (5649), 1380.

26 Organick, L., *et al*. Random access in large-scale DNA data storage, *Nat. Biotechnol.*, In Press.

27 Carlson, R. (2017) Guesstimating the Size of the Global Array Synthesis Market, http://www.synthesis.cc/synthesis/2017/8/guesstimating-the-size-of-the-global-array-synthesis-market (accessed 3 January 2018).

28 National Security Council (2009) The National Strategy for Countering Biological Threats.

29 Goldberg, M. (2013) BioFab: applying Moore's Law to DNA synthesis. *Ind. Biotechnol.*, **9** (1), 10–12.

30 Carlson (2010).

31 Carlson, R. (2016) Estimating the biotech sector's contribution to the US economy. *Nat. Biotechnol.*, **34**, 247–255.

2

Trackable Multiplex Recombineering (TRMR) and Next-Generation Genome Design Technologies: Modifying Gene Expression in *E. coli* by Inserting Synthetic DNA Cassettes and Molecular Barcodes

Emily F. Freed[1], Gur Pines[2,3], Carrie A. Eckert[1,3], and Ryan T. Gill[2,3]

[1] *Biosciences Center, National Renewable Energy Laboratory, 15013 Denver West Parkway, Golden, CO 80401, USA*
[2] *University of Colorado, Chemical and Biological Engineering, 3415 Colorado Ave, Boulder, CO, 80303 USA*
[3] *University of Colorado, Renewable and Sustainable Energy Institute, 4001 Discovery Dr, Boulder, CO 80303 USA*

Trackable multiplex recombineering (TRMR) allows researchers to explore the otherwise large mutational space of the *Escherichia coli* genome efficiently. This method is used to simultaneously change the expression level of every gene in the genome, so that each gene is either overexpressed or switched off. A variation on TRMR, tunable trackable multiplex recombineering (T^2RMR), allows expression levels to be tuned over a 10^4-fold range. TRMR and T^2RMR therefore allow bacterial responses to be tuned to different environmental cues. Additionally, the genomic changes can be tracked and identified for population dynamic studies and for further analyses thanks to "barcoding" (or "tagging") of every mutation. The TRMR and T^2RMR procedures include library design, production, and amplification, followed by the insertion of the DNA library into a precise location in the genome via phage-enabled homologous recombination. Then, the heterogeneous bacterial population is subjected to a defined stress or screened for a specific trait. Finally, beneficial mutations are identified by means of barcode hybridization to a microarray or by sequencing. Importantly, TRMR- and T^2RMR-based populations can be established by a single scientist in a single day, and depending on the desired trait, genome-wide mapping results may be obtained as shortly as within a week.

2.1 Introduction

While traditional engineering usually involves the design and production of mechanical structures and devices, biological engineering is focused on modifying the natural world and adapting it to human needs. Although many consider biological engineering to be a new field, it is as old as civilization itself. Selective breeding of plants and animals for specific traits exemplifies one of the hallmarks of biological engineering and evolution in general: selection of a successful subpopulation that will establish the next generations, thus continuously refining

Synthetic Biology: Parts, Devices and Applications, First Edition. Edited by Christina Smolke.

the desired phenotype. The field of genetics and the discovery of the mechanisms by which traits are propagated along generations allowed people, for the first time, to rationally induce genetic modifications rather than wait for them to occur randomly. Early efforts focused on rational transfer and modifications of single genes and were collectively termed "genetic engineering." Complex traits, however, derived from multiple gene interactions or whole metabolic pathways, cannot be efficiently engineered one gene at a time and require high-throughput and systemic approaches, the recognition of which formed the basis of the field of metabolic engineering [1, 2].

Since then, advances in DNA sequencing and systems biology methodologies have led to exceptional new approaches for characterizing complex traits and their underlying genetic networks. Additionally, rapid progress in DNA chemical synthesis and the development of recombination-based methods now allow mutations to be incorporated in multiplex and at a throughput orders of magnitudes beyond the state of the art a decade ago [3–6]. In contrast to earlier methods of individually synthesizing oligonucleotides or using DNA segments from natural sources, current technology allows the parallel production of synthetic DNA (synDNA) libraries [5]. Additionally, homologous recombination-based techniques (recombineering) that promote the integration of foreign DNA into the chromosome of the target organism have reached relatively high levels of efficiency [6]. Recombineering in *E. coli* is based on targeting a synthetic recombineering substrate (a single-stranded (ss) DNA oligonucleotide or a double-stranded (ds) DNA cassette) to a specific locus on the chromosome via homology arms. Typically, this DNA substrate contains a desired mutation and may also code for an antibiotic resistance gene as a selective marker. The actual recombination is enabled by either the Rec E/T or the λ-Red prophage system [6, 7].

Here, we describe the TRMR and T^2RMR techniques, which not only make the multiplexing of recombineering possible in *E. coli* but also provide the ability to track the engineered genetic changes accurately. Currently, both library designs allow one to target, in parallel, every gene in the genome for either overexpression or downregulation, with T^2RMR allowing for tuning of gene expression over a $\sim 10^4$-fold range. The trackability is achieved by adding a unique "molecular barcode" [8] upstream of every mutation, facilitating its identification. These methods enable the search for specific and desired genetic traits and aid in the navigation of an otherwise large mutational space (i.e., in this case, the total number of possible single mutations). We discuss the benefits of such methods, existing challenges, possible combinations with other methods, and some possible future development and applications.

2.2 Current Recombineering Techniques

E. coli did not evolve an efficient mechanism for recombination; therefore spontaneous homologous recombination of foreign genetic material is typically a rare event, on the order of 10^{-6} for linear ssDNA or dsDNA substrates [9]. It has been suggested that the low efficiency is primarily due to endogenous nucleases that rapidly degrade the unprotected DNA [10, 11]. Phage-based methods, which rely

on the overexpression of phage proteins that prepare and protect the DNA substrate, have been developed to more efficiently induce the incorporation of desired DNA segments into cells. Two popular methods are the Rec E/T and the λ-Red prophage systems.

2.2.1 Recombineering Systems

The Rec E/T system encodes for two proteins, namely, RecE, a 5′ → 3′ exonuclease, and RecT, an ssDNA binding protein [7]. The λ-Red system encodes for three proteins: Exo (homologous to RecE), Beta (homologous to RecT), and Gam, an inhibitor of the endogenous RecBCD exonuclease, which acts to protect the foreign DNA from active degradation [12]. The foreign DNA to be recombineered into the host genome may be in one of two forms depending on its source. Synthetic oligonucleotides (oligos) will usually be ssDNA, while polymerase chain reaction (PCR)-amplified segments are double stranded. In order to be incorporated into the host genome, the recombineering substrate includes the desired DNA sequence to be incorporated and homology regions that flank both sides of this DNA sequence. The homology regions direct the DNA substrate to a specific location in the genome, where the endogenous replication machinery uses it as a template for replication. Here, we will focus on the λ-Red system, in which the Exo, Beta, and Gam proteins work in concert to induce homologous recombination of DNA fragments into the host genome.

Currently, the most popular vectors for the λ-Red system are either the pKD46 or the pSIM5 plasmids [11, 13, 14]. Protein expression from these vectors is induced by incubation with arabinose or at 42 °C, respectively. Both are additionally temperature sensitive at 37 °C, which allows for plasmid curing following expression of the λ-Red proteins. The standard λ-Red recombineering workflow includes transforming the host strain with the recombineering plasmid of choice, induction of the recombineering machinery, and additional transformation with the desired recombineering substrate, followed by selection/screening for successful recombinant strains [11].

2.2.2 Current Model of Recombination

Several models of the exact recombination mechanism exist; however, the "replication fork annealing model" is currently the most supported experimentally (Figure 2.1). According to this model, if the recombineering substrate consists of dsDNA, the λ-Red Exo protein, through its exonuclease activity, transforms it into ssDNA [16]. This model suggests that although some dsDNA is being completely degraded by Exo molecules that digest the recombineering substrate from both sides, in some cases one strand is digested completely before the other side is attacked by another Exo molecule, rendering the resulting ssDNA immune from further Exo digestion. If the recombineering substrate is ssDNA, no action by Exo is required. In both cases, the (resulting) ssDNA strand is protected from further degradation by endogenous nucleases via Beta, which binds to the ssDNA and escorts it to single-stranded areas in the chromosome [17, 18]. Single-stranded regions occur during DNA repair, transcription-induced supercoiling,

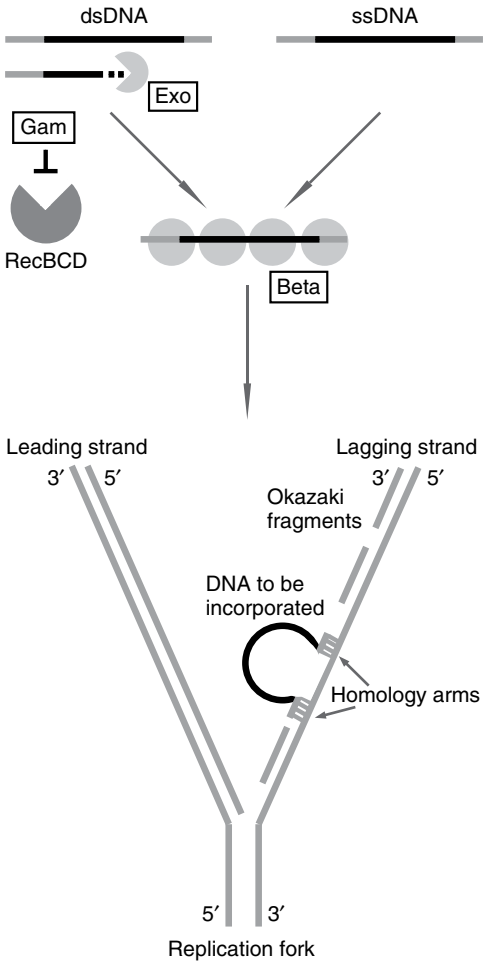


Figure 2.1 The λ-Red system and the replication fork annealing model of recombination. Either a double- or single-stranded recombineering substrate, consisting of the DNA sequence to be inserted flanked by homology arms, is transformed into cells. The λ-Red proteins facilitate recombination by digesting one strand of DNA in the case of dsDNA (Exo), by inhibiting RecBCD nuclease activity (Gam), and by protecting and conveying the ssDNA to the replication fork (Beta). Then, the ssDNA acts as a mismatched Okazaki fragment and binds to the lagging strand via its homology arms. This process results, upon completion of cell duplication, with one wild-type daughter cell and one recombineered, heterozygous-like daughter cell. Reprinted with permission from Pines *et al.* 2015 [15]. Copyright 2015 American Chemical Society.

and, importantly, all along the genome during chromosome replication. Studies show that the ssDNA substrate anneals to the chromosome in a strand-biased manner, which correlates with the direction of DNA replication [19, 20]. These results suggest that the ssDNA annealing is directed to the lagging strand of the replication fork via its homology regions, where it essentially acts as an exogenous DNA-based Okazaki fragment [16, 21]. Overall, the recombineering

process results in two daughter cells, one of which harbors the desired genetic modification, while the other remains genetically identical to its parental ancestor, limiting this method to a theoretical maximum efficiency of 50% [15].

2.3 Trackable Multiplex Recombineering

The *E. coli* genome consists of over 4000 genes. When engineering the *E. coli* genome for a desired trait (e.g., tolerance to a growth condition or increased production of a valuable chemical), combinations of multiple genetic modifications are often required to achieve optimal performance. The result is a combinatorial mutation space that expands exponentially with the number of targeted genes and quickly exceeds the size of space that can be searched on laboratory time scales. For example, if each of the 4000 genes is modified to both an "off" and an "on" state, there are 2^{4000} possible states. TRMR and T^2RMR provide a rapid and efficient way to modify an entire genome in a controlled manner and to evaluate the effects of those genetic modifications simultaneously. Using these techniques it is possible to modify >95% of the genes in *E. coli* in a single day. An overview of the TRMR and T^2RMR techniques is shown in Figure 2.2. In order to engineer a genome using TRMR or T^2RMR, a synDNA cassette is created that encodes for a genetic feature (such as the overexpression or underexpression of each specific gene) and a molecular barcode that is used to track each feature. These synDNA cassettes are then introduced in parallel into cell populations via recombineering. Next, the modified populations are grown in any desired growth condition or in selective medium. Microarray or sequencing analysis of the molecular barcodes is used to determine the relative fitness of each allele/engineered cell in the surviving population under the chosen conditions. TRMR and T^2RMR libraries must be used to evaluate a phenotype that can be either selected or screened for.

To date, TRMR and T^2RMR have been used to map genes required for growth in various types of media and to optimize tolerance to acetate, low pH, cellulosic hydrolysate, isobutanol, ethanol, isopentenol, furfural, and various antibiotics [22–26]. These studies have given insight into carbon source and vitamin utilization, primary and secondary metabolism, and mechanisms of toxicity under a variety of conditions.

While the next few paragraphs will provide basic information on the TRMR and T^2RMR methods, readers are referred to an in-depth protocol for complete experimental details of TRMR [27].

2.3.1 TRMR and T^2RMR Library Design and Construction

All TRMR and T^2RMR libraries consist of two main parts: (i) "targeting" oligos that contain homology to each gene in the *E. coli* genome, a molecular barcode to identify each oligo uniquely, and sequences used to amplify each region of the oligo by PCR, and (ii) "shared DNA" that encodes for a genetic function and an antibiotic resistance marker. These two parts are then ligated to each other to create synDNA cassettes that can then be amplified, linearized, and transformed

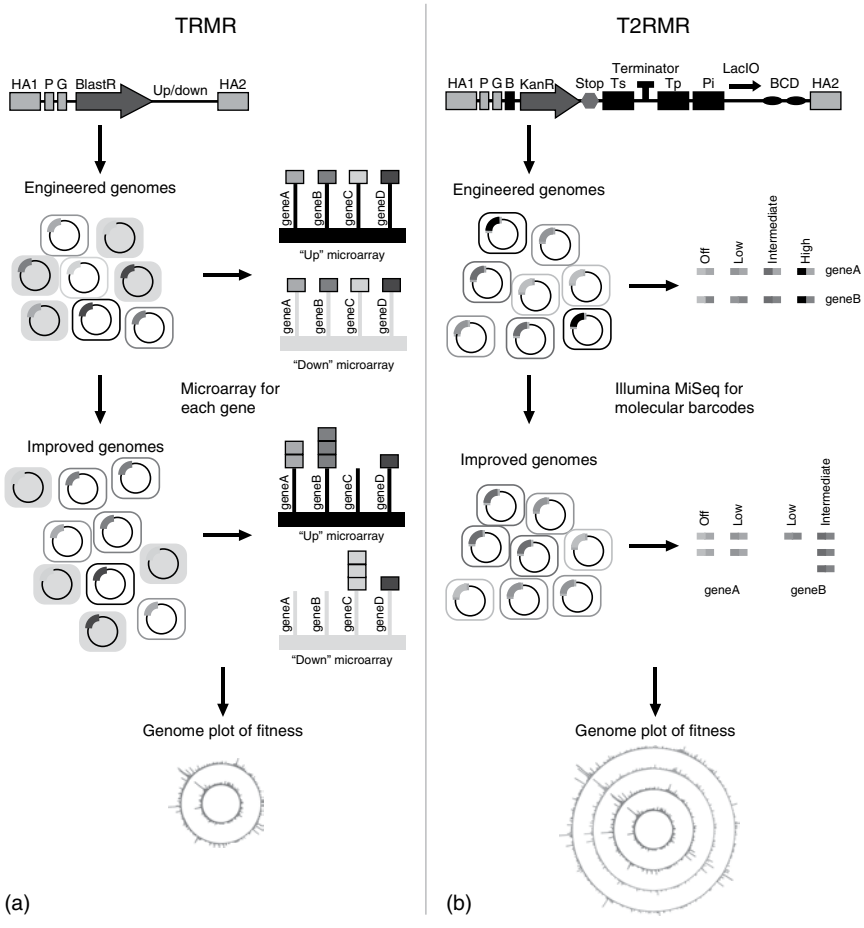


Figure 2.2 Overview of trackable multiplex recombineering (a) and tunable trackable multiplex recombineering (b). TRMR and T^2RMR cassettes are designed and synthesized in multiplex followed by transformation into *Escherichia coli*. The *E. coli* population is then placed under selective pressure. Alleles that are enriched during selection are identified by microarray (TRMR) or next-generation sequencing (T^2RMR), and their relative fitness is determined. Cassette design for each technique is shown at the top. Black regions are shared DNA and gray regions are from the targeting oligos. HA1 and HA2, homology regions; P, barcode priming site; G, barcode identifying the gene; B, barcode identifying the BCD; BlastR, blasticidin resistance gene; KanR, kanamycin resistance gene; stop, three frame stop codons; Ts, terminator spacer; Tp, terminator pause; Pi, promoter insulator; LacIO, LacI-regulated synthetic inducible promoter (apFAB906); BCD, bicistronic design (dual RBS). Adapted with permission from Freed *et al.* 2015 [22]. Copyright 2015 American Chemical Society.

into cells (Figure 2.3). During amplification, circular concatemers are created. These concatemers are then cleaved by restriction enzyme digest to generate linear dsDNA with the homology regions, molecular barcodes, antibiotic resistance, and gene modification sequences in the correct order. The synDNA cassettes are linearized because linear DNA, and not circular DNA, is generally used as a substrate for recombineering using the λ-Red system [6].

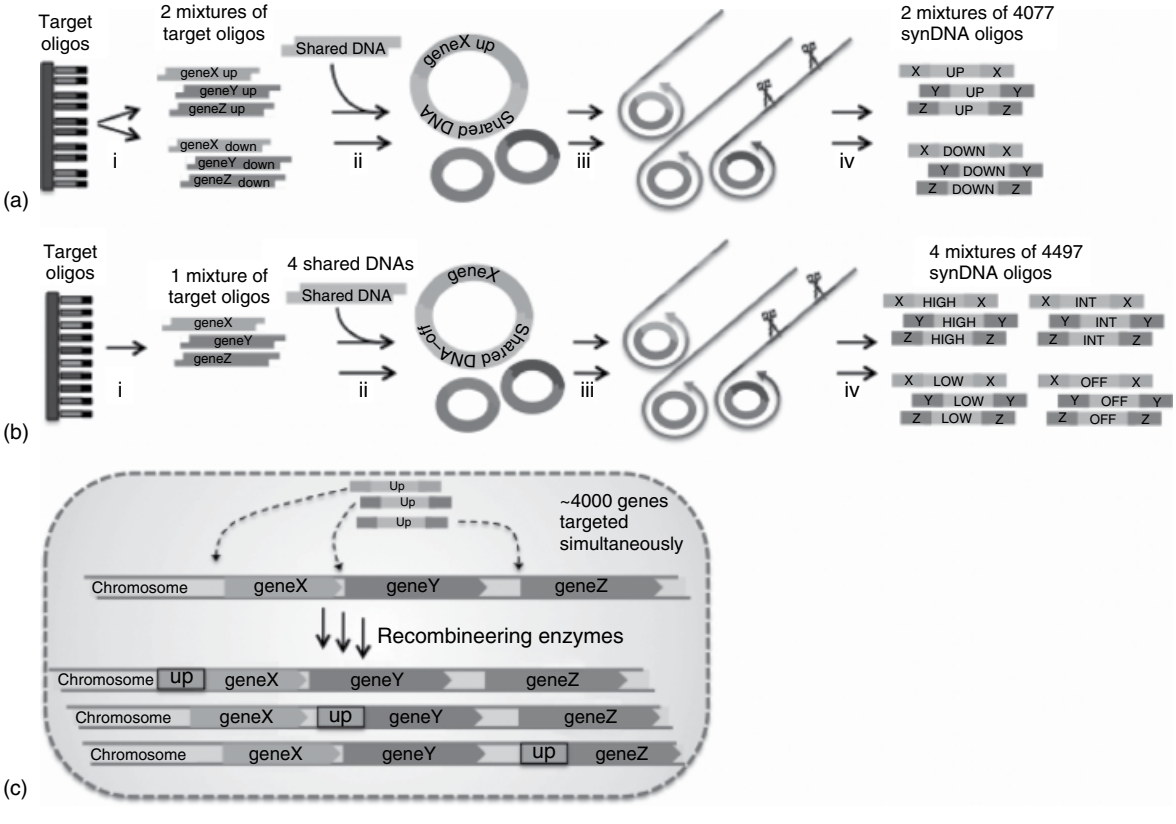


Figure 2.3 Construction and incorporation of TRMR (a) and T^2RMR (b) cassettes. In both TRMR and T^2RMR, targeting oligos are ligated with a shared DNA cassette encoding a specific genetic function. The ligated synDNA cassettes are amplified by rolling circle amplification and then cleaved to create a linear dsDNA substrate. (c) The linear synDNA cassettes are recombineered into cells, targeting all genes at one time. Adapted from Warner *et al.* 2010 [26].

In both TRMR and T^{2}RMR, targeting oligos were designed using homology regions that result in the synDNA cassette being inserted upstream of each gene and replacing each gene's start codon in *E. coli* MG1655. In the original demonstration of TRMR [26], all protein-coding genes were targeted. Each targeting oligo also contained a unique 20-nucleotide sequence that served as a molecular barcode (or "tag") used to track each gene. The barcodes were chosen from a set that had previously been used successfully in yeast [8]. In T^{2}RMR [22], pseudogenes and noncoding RNAs were targeted in addition to all protein-coding genes. Each targeting oligo contained a 12-nucleotide sequence that had been optimized to serve as a molecular barcode for high-throughput sequencing. Targeting oligos for both TRMR and T^{2}RMR were synthesized on a microchip by Agilent.

The most significant differences between TRMR and T^{2}RMR are in the design of the shared DNA. TRMR consists of two libraries: an "up" library that causes genes to be overexpressed and a "down" library that causes genes to be underexpressed. The shared DNA for the up cassette contains the strong PLtetO-1 promoter and a strong ribosome binding site (RBS), which generally results in increased transcription and translation of downstream genes. The shared DNA for the down cassette contains no promoter and no RBS, resulting in the deletion of the native RBS and subsequent decrease in translation initiation. The activity of the β-galactosidase protein (*lacZ* gene) was used to confirm that the up construct for this protein resulted in overexpression and the down construct resulted in loss of expression of the protein (Figure 2.4a).

While these libraries have been successful in identifying alleles responsible for a desired phenotype, there are some limitations to the original TRMR libraries. One drawback is that these libraries do not use standardized synthetic parts, which may result in inconsistent expression levels across targeted genes, since it is known that placing the same promoter or RBS in front of two different genes can cause the two genes to be expressed at vastly different levels [22–25]. Another

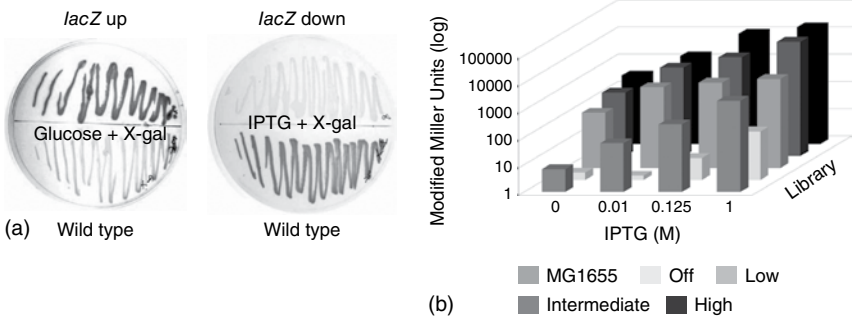


Figure 2.4 Validation of TRMR (a) and T^{2}RMR (b) cassettes using the *lacZ* gene. In TRMR, the "up" cassette causes *lacZ* to be expressed, while the "down" cassette results in a loss of expression. In T^{2}RMR, varying the four libraries ("off," "low," "intermediate," and "high"), and the amount of inducer (IPTG) allows *lacZ* to be expressed over a ~10^4-fold range. Adapted with permission from Warner *et al.* 2010 [26] and Freed *et al.* 2015[22]. Copyright 2015 American Chemical Society.

limitation is that the original TRMR libraries are binary in nature and do not allow the tuning of expression levels. When engineering a synthetic pathway, the activity of the pathway can be dependent on expressing each protein in a very narrow expression range [28] rather than in a binary way.

The design of the shared DNA for T^{2}RMR addresses these limitations. T^{2}RMR uses a "bicistronic design" (BCD), devised by Mutalik *et al.* [29], that embeds two RBSs in a nucleotide sequence that encodes a 16-amino-acid peptide, which is then placed directly in front of a gene of interest. The dual RBSs embedded in this leader peptide affect ribosome binding and translation initiation, and therefore these BCD constructs give much more consistent expression when tested with a variety of genes. Mutalik *et al.* further tested several hundred promoter variants, as well as combinations of BCDs and promoters, to express genes over a wide dynamic range. In 93% of cases, they found they could predict the expression level of a protein to within twofold, which is a great improvement over using a single RBS [29]. T^{2}RMR consists of four libraries/shared DNAs, each expressing a different BCD, to give four base expression levels: "off," "low," "intermediate," and "high." Each library contains a 12-nucleotide barcode to identify the BCD during high-throughput sequencing. An inducible LacI-regulated promoter was placed in front of each BCD allowing for fine-tuning of gene expression to almost any level that is desired just by changing the amount of inducer (isopropyl β-D-1-thiogalactopyranoside (IPTG)) that is added to the medium. Validation of T^{2}RMR with the β-galactosidase protein (*lacZ* gene) gave expression over a ~10^4-fold activity range, as measured by the Miller assay, by using different combinations of the libraries and amounts of IPTG (Figure 2.4b).

2.3.2 Experimental Procedure

The experimental procedure is the same for both TRMR and T^{2}RMR. Once the double-stranded, linearized synDNA libraries have been constructed, they are incorporated into the genome by homologous recombination. *E. coli* cells containing the λ-Red recombination proteins (either integrated directly on the chromosome or expressed from a plasmid such as pSIM5 [14]) are grown at 30 °C to mid-log phase in medium containing the appropriate antibiotic if required. If using pSIM5, expression of the λ-Red enzymes is induced by incubating cells at 42 °C for 15 min. The cells are then chilled on ice and made electrocompetent by washing with ice cold water as previously described [11]. SynDNA is then transformed into cells by electroporation. After 2 h of recovery at 37 °C, cells are spread on plates containing the antibiotic resistance marker that is selective for library clones. Plates are incubated at 37 °C for 22 h and then colonies are scraped from the plates, resuspended in LB medium, and aliquoted for storage at −70 °C.

Screening and selection of TRMR and T^{2}RMR clones can be carried out using either liquid or solid media with any chemical compound that modifies growth or confers a phenotype that can be detected by a high-throughput assay. An equal number of cells from all libraries (either up and down or off, low, intermediate, and high) are mixed together in medium containing the antibiotic that is

selective for the libraries and are grown to late log phase. An aliquot of this culture is frozen for further analysis, with the remainder of the culture being used for selections. For selections in liquid medium, cells from the initial culture are inoculated into medium containing the desired chemical compound and are grown to stationary phase. Cells are then harvested for analysis as discussed in the following section. For selections on solid medium, cells from the initial culture are spread on plates containing the desired chemical compound, and plates are incubated until colonies are visible. All colonies are scraped from the plates for further analysis.

2.3.3 Analysis of Results

Either microarray or sequencing analysis can be used to determine the relative fitness of each allele after selection. Genomic DNA is extracted from cells before and after selection (and at various points during selection if desired), and the molecular barcodes from each sample are amplified by PCR. In the original set of TRMR experiments, barcodes were hybridized to the GenFlex Tag4 array from Affymetrix [30]. A separate microarray experiment was done for each library. Ten barcodes were spiked into the genomic DNA mixture in known amounts to determine barcode concentrations, and 1642 negative-control barcodes were also included to determine background hybridization rates. Allele frequencies for each gene were then determined by dividing allele concentrations by the total concentration of all alleles detected on the array.

In T^2RMR, molecular barcodes that are optimized for high-throughput sequencing were used instead of microarray to track alleles. Each allele had two barcodes – one identifying the library (off, low, intermediate, high) and one identifying the gene. All samples were combined into a single MiSeq run. High-throughput sequencing allows more quantitative analysis of genotype frequencies, since individual alleles are directly tracked at the nucleotide level rather than by relative hybridization intensity (measured in arbitrary fluorescence units). A single run of Illumina MiSeq can generate 10^6–10^7 sequencing reads (a typical microarray signal distribution ranges over about 10^3), allowing for each barcode to be sequenced thousands of times. This deep sequencing additionally aids in the detection of rare alleles [31], which might be present in too low a concentration to be identified by microarray hybridization. A microarray hybridization signal can also saturate [30], resulting in loss of data for the most highly expressed alleles. A second advantage to high-throughput sequencing is that it results in a lower error rate in identifying alleles. Although hybridization to a microarray, in general, gives high fidelity, some alleles will fail to hybridize to the sequence on the microarray that is perfectly complementary [32]. Furthermore, errors in barcode sequences can be introduced during cell replication, DNA synthesis, or PCR amplification, resulting in loss of hybridization or, worse, in hybridization to an incorrect spot on the microarray [32]. With high-throughput sequencing, on the other hand, errors in barcode sequences can be identified and corrected or discarded [33], as was done with T^2RMR. In T^2RMR, allele frequencies for each gene were determined by dividing the number of barcode counts for that gene by the total number of barcode counts for all genes.

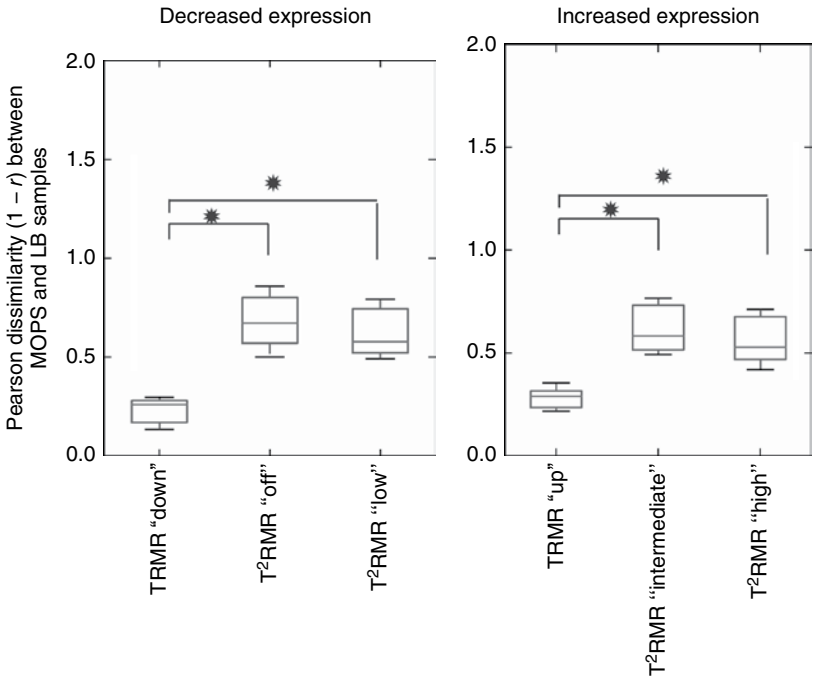


Figure 2.5 T^2RMR has significantly increased ability to discriminate between MOPS minimal medium and LB-rich medium. The Pearson dissimilarity (0 indicates perfectly linearly correlated, and 2 indicates negatively correlated) between MOPS and LB samples for each library type is shown. * indicates $p < 0.05$ Benjamini–Hochberg corrected significance. Adapted with permission from Freed *et al.* 2015 [22]. Copyright 2015 American Chemical Society.

In both TRMR and T^2RMR, relative fitness was calculated by determining the ratio of the final allele frequency after selection to the initial allele frequency. Any allele that increases in frequency after selection is likely to confer tolerance to the selective condition. While both TRMR and T^2RMR are successful at identifying alleles responsible for a desired phenotype, T^2RMR may be able to identify alleles that improve fitness under weak selective pressure that the original TRMR would not be able to identify. For example, T^2RMR does significantly better than TRMR at discriminating between LB and MOPS growth media (Figure 2.5). In all cases, to confirm a fitness advantage, it is advisable to analyze the growth of cells containing each individual allele that is enriched during selection and compare it with wild-type cells.

2.4 Current Challenges

TRMR and T^2RMR are novel and powerful techniques that allow for the modification and tracking of thousands of genes in a single step. However, there are some challenges that need to be considered when performing TRMR or T^2RMR.

2.4.1 TRMR and T^{2}RMR are Currently Not Recursive

One challenge for the current TRMR and T^{2}RMR designs is that only a single round of recombineering is currently implemented. This limitation is due to the fact that the TRMR and T^{2}RMR substrates are dsDNA, and because of the current low efficiency of dsDNA recombineering, it is essential that an antibiotic selection step is used to ensure the removal of non-recombineered cells. While multiple different antibiotic markers may be used for successive rounds of recombination with a TRMR or T^{2}RMR library, the limited number of markers restricts the number of additional cycles that can be performed. Another option is to remove the resistance gene (via flanking FRT sites) and reintroduce the same library in the next recombineering round, but this will greatly extend the time required for every cycle. Alternately, a new technology called CRISPR-enabled trackable genome engineering (CREATE) has been developed that allows clustered regularly interspaced short palindromic repeat (CRISPR)-based markerless selection of recombineered cells [34]; this technique could be combined with the expression-level cassettes from TRMR or T^{2}RMR as discussed in Section 2.5.2.

An additional concern is the rapid increase in the number of recombinants that result from every TRMR or T^{2}RMR cycle. In the ideal case, every possible combination of mutations should be represented in the cell population, which would require a volume of cells that exceeds the capability of current equipment. Lastly, barcode identification is being performed at the whole population level and thus does not distinguish whether the barcode originated from a single or multiple cells. This limitation could be overcome by using new single-cell sequencing technologies including (i) single-cell linkage PCR, which allows for the sequencing of millions of barcoded individual cells [35] and (ii) tracking combinatorial engineered libraries (TRACE), which gives the ability to track combinations of mutations from a single cell [36, 37].

2.4.2 Need for More Predictable Models

Mathematical modeling of a metabolic pathway can be a valuable tool for further optimization of that pathway [reviewed in [38]]. Once the metabolic flux through a pathway is accurately modeled, bottlenecks in that pathway can be identified, and further engineering efforts can be directed toward removing that bottleneck. TRMR and T^{2}RMR can aid in the development of models by identifying genes that are involved in a pathway and that would have been difficult to predict *a priori* [22, 26]. Once these genes have been identified, new TRMR-like libraries that are predicted to be enriched for better performing strains can be designed.

Although metabolic models can be useful, unfortunately they often lack predictive power [reviewed in [39]]. This can be due to a number of factors including lack of mechanistic detail about the pathway, inconsistent behavior of synDNA parts, or failure to account for epistatic interactions [25]. Epistatic interactions can make both TRMR and T^{2}RMR data particularly difficult to model. The development of more predictive models is an active and ongoing part of metabolic engineering research.

2.5 Complementing Technologies

2.5.1 MAGE

The TRMR and T^2RMR approach of targeting all genes simultaneously in a trackable manner may prove beneficial for selecting candidate genes for other downstream techniques in the pursuit of improved production of chemicals and for strain development. The advantage TRMR and T^2RMR provide is the potential discovery of genes whose involvement in any specific pathway is currently impossible to predict using computational or other methods. For example, gene candidates can be derived from a tolerance experiment, as mentioned earlier. However tolerance may be increased even further by combining several such mutations via multiplex automated genome engineering (MAGE) [40] or by using another combinatorial, recursive multiplex recombineering technique. Not only do such combinations dramatically increase the mutational space, but combinatorial experiments also must consider epistatic effects among the combined mutations, which are extremely difficult to predict *a priori* [25]. Additionally, the question of which candidate genes should be included in the second-step MAGE experiment is far from trivial. Intuitively, one might pick the top performing genes for combinatorial experiments. However this approach, termed the "greedy approach," might result in reaching a local maximum in the potential fitness landscape rather than the desired global maximum. Current computational efforts are being carried out to tackle these challenges.

2.5.2 CREATE

To date, TRMR and T^2RMR have only been used for modifying the expression level of genes. However, work done on the engineering of biocatalysts has shown that in some cases a single point mutation can alter the catalytic activity of an enzyme (reviewed in [41]), suggesting that big advances can come from subtle changes. Barcoded editing at the single nucleotide polymorphism (SNP) level will therefore lead to even faster improvements in strain engineering and pathway optimization.

Recent technologies were designed to address this need for higher-resolution genome editing, namely, creating point mutations within an open reading frame. The first generation of these ideas took advantage of the newly discovered CRISPR/Cas9 systems to increase editing efficiency and introduce single edits [42–44]. Multiplex editing of numerous sites quickly followed [34]. Similar to TRMR and T^2RMR, CREATE utilizes the λ-Red recombineering system and array-based DNA synthesis to create rationally designed edits. Here, however, the CRISPR/Cas9 system is used for increasing editing efficiency and for the removal of non-edited genomes. CRISPR/Cas9-based editing technologies take advantage of the RNA-guided endonuclease activity of the Cas9 protein [45]. This activity depends on a dinucleotide GG protospacer adjacent motif (PAM), leading to a site-specific double-strand break in the cell's genomic DNA and subsequent cell death (in cells deficient of an efficient double-strand break repair

mechanism, such as *E. coli*). However, cells containing a mutation in the PAM site are protected from DNA cleavage [42].

Each CREATE cassette includes the target site in a gene and a proximal PAM sequence that are selected for mutagenesis. Since in most cases the PAM falls within the open reading frame, it is silently mutated, so the amino acid sequence is not altered. To allow multiplex editing, the PAM-specific corresponding gRNA coding sequence is co-synthesized with the target site editing oligo and cloned into an editing vector, which also serves as a target-specific barcode. This design enables the creation of barcoded libraries composed of tens of thousands of cells, with each genome having a single amino acid edit. Hence, TRMR or T^2RMR and CREATE can be used in conjunction, with TRMR or T^2RMR identifying important genes under specific conditions and CREATE allowing for the engineering of those genes for optimal results. These two technologies could additionally be combined in the future (this would require new technology allowing for an increase in the length of targeting oligos that can be synthesized on a microchip). Adding the main T^2RMR elements to the CREATE cassette design can allow higher versatility in editing. Gene expression tuning can be coupled to gene editing, enabling the investigation of expression in conjunction with point mutations. The ability to cycle these edits for the generation of multiple diverse genotypes will help researchers to isolate desired complex traits that combine both protein sequence and expression level.

2.6 Conclusions

Recent advances in DNA synthesis and the development of standardized genetic parts have greatly increased genome engineering capabilities. TRMR and T^2RMR allow a single researcher to modify an entire genome in a single day and map which alleles are responsible for a desired phenotype. This ability to fine-tune expression levels, particularly when combined with other technologies for making point mutations or combinatorial mutations, will allow researchers to quickly and easily engineer strains for maximal production of, or tolerance to, any compound.

Definitions

Recombineering Genetic engineering using homologous recombination

TRMR Trackable multiplex recombineering

Recombineering substrate A single- or double-stranded piece of DNA that is to be inserted into the target's genome via recombineering

λ-Red system Using λ-Red phage proteins to enable homologous recombination in bacteria

Molecular barcode Unique, short nucleotide sequence used to identify and track a specific gene or piece of DNA

Synthetic DNA Artificial DNA that is synthesized; does not have to be based on naturally occurring DNA sequence

References

1 Bailey, J.E. (1991) Toward a science of metabolic engineering. *Science*, **252** (5013), 1668–1675.
2 Stephanopoulos, G. and Vallino, J.J. (1991) Network rigidity and metabolic engineering in metabolite overproduction. *Science*, **252** (5013), 1675–1681.
3 Shendure, J. and Lieberman Aiden, E. (2012) The expanding scope of DNA sequencing. *Nat. Biotechnol.*, **30** (11), 1084–1094.
4 Oberhardt, M.A., Palsson, B.Ø., and Papin, J.A. (2009) Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.*, **5** (1), 320.
5 Cleary, M.A., Kilian, K., Wang, Y. *et al.* (2004) Production of complex nucleic acid libraries using highly parallel in situ oligonucleotide synthesis. *Nat. Methods*, **1** (3), 241–248.
6 Yu, D., Ellis, H.M., Lee, E.-C. *et al.* (2000) An efficient recombination system for chromosome engineering in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.*, **97** (11), 5978–5983.
7 Zhang, Y., Buchholz, F., Muyrers, J.P.P., and Stewart, A.F. (1998) A new logic for DNA engineering using recombination in *Escherichia coli*. *Nat. Genet.*, **20** (2), 123–128.
8 Shoemaker, D.D., Lashkari, D.A., Morris, D. *et al.* (1996) Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar–coding strategy. *Nat. Genet.*, **14** (4), 450–456.
9 Murphy, K.C. (1998) Use of bacteriophage λ recombination functions to promote gene replacement in *Escherichia coli*. *J. Bacteriol.*, **180** (8), 2063–2071.
10 Datta, S., Costantino, N., Zhou, X., and Court, D.L. (2008) Identification and analysis of recombineering functions from gram-negative and gram-positive bacteria and their phages. *Proc. Natl. Acad. Sci. U.S.A.*, **105** (5), 1626–1631.
11 Sharan, S.K., Thomason, L.C., Kuznetsov, S.G., and Court, D.L. (2009) Recombineering: a homologous recombination-based method of genetic engineering. *Nat. Protoc.*, **4** (2), 206–223.
12 Murphy, K.C. (1991) Lambda Gam protein inhibits the helicase and chi-stimulated recombination activities of *Escherichia coli* RecBCD enzyme. *J. Bacteriol.*, **173** (18), 5808–5821.
13 Cotta-de-Almeida, V., Schonhoff, S., Shibata, T. *et al.* (2003) A new method for rapidly generating gene-targeting vectors by engineering BACs through homologous recombination in bacteria. *Genome Res.*, **13** (9), 2190–2194.
14 Datta, S., Costantino, N., and Court, D.L. (2006) A set of recombineering plasmids for gram-negative bacteria. *Gene*, **379**, 109–115.
15 Pines, G., Freed, E.F., Winkler, J.D., and Gill, R.T. (2015) Bacterial recombineering: genome engineering via phage-based homologous recombination. *ACS Synth. Biol.*, **4** (11), 1176–1185.
16 Mosberg, J.A., Lajoie, M.J., and Church, G.M. (2010) Lambda red recombineering in *Escherichia coli* occurs through a fully single-stranded intermediate. *Genetics*, **186** (3), 791–799.
17 Muniyappa, K. and Radding, C.M. (1986) The homologous recombination system of phage lambda. Pairing activities of beta protein. *J. Biol. Chem.*, **261** (16), 7472–7478.

18 Karakousis, G., Ye, N., Li, Z. *et al.* (1998) The beta protein of phage λ binds preferentially to an intermediate in DNA renaturation. *J. Mol. Biol.*, **276** (4), 721–731.

19 Ellis, H.M., Yu, D., DiTizio, T., and Court, D.L. (2001) High efficiency mutagenesis, repair, and engineering of chromosomal DNA using single-stranded oligonucleotides. *Proc. Natl. Acad. Sci. U.S.A.*, **98** (12), 6742–6746.

20 Lim, S.I., Min, B.E., and Jung, G.Y. (2008) Lagging strand-biased initiation of red recombination by linear double-stranded DNAs. *J. Mol. Biol.*, **384** (5), 1098–1105.

21 Maresca, M., Erler, A., Fu, J. *et al.* (2010) Single-stranded heteroduplex intermediates in λ red homologous recombination. *BMC Mol. Biol.*, **11** (1), 54.

22 Freed, E.F., Winkler, J.D., Weiss, S.J. *et al.* (2015) Genome-wide tuning of protein expression levels to rapidly engineer microbial traits. *ACS Synth. Biol.*, **4** (11), 1244–1253.

23 Weiss, S.J., Mansell, T.J., Mortazavi *et al.* (2016) Parallel mapping of antibiotic resistance alleles in *Escherichia coli*. *PLoS One*, **11** (1), e0146916.

24 Glebes, T.Y., Sandoval, N.R., and Gillis, J.H. (2015) Comparison of genome-wide selection strategies to identify furfural tolerance genes in *Escherichia coli*. *Biotechnol. Bioeng.*, **112**, 129–140.

25 Sandoval, N.R., Kim, J.Y.H., Glebes, T.Y. *et al.* (2012) Strategy for directing combinatorial genome engineering in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.*, **109** (26), 10540–10545.

26 Warner, J.R., Reeder, P.J., Karimpour-Fard, A. *et al.* (2010) Rapid profiling of a microbial genome using mixtures of barcoded oligonucleotides. *Nat. Biotechnol.*, **28** (8), 856–862.

27 Mansell, T.J., Warner, J.R., and Gill, R.T. (2013) Trackable multiplex recombineering for gene-trait mapping in *E. coli*. *Methods Mol. Biol.*, **985**, 223–246.

28 Temme, K., Zhao, D., and Voigt, C.A. (2012) Refactoring the nitrogen fixation gene cluster from *Klebsiella oxytoca*. *Proc. Natl. Acad. Sci. U.S.A.*, **109** (18), 7085–7090.

29 Mutalik, V.K., Guimaraes, J.C., Cambray, G. *et al.* (2013) Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods*, **10** (4), 354–360.

30 Pierce, S.E., Fung, E.L., Jaramillo, D.F. *et al.* (2006) A unique and universal molecular barcode array. *Nat. Methods*, **3** (8), 601–603.

31 Robins, W.P., Faruque, S.M., and Mekalanos, J.J. (2013) Coupling mutagenesis and parallel deep sequencing to probe essential residues in a genome or gene. *Proc. Natl. Acad. Sci. U.S.A.*, **110** (9), E848–E857.

32 Cho, R.J., Fromont-Racine, M., Wodicka, L. *et al.* (1998) Parallel analysis of genetic selections using whole genome oligonucleotide arrays. *Proc. Natl. Acad. Sci. U.S.A.*, **95** (7), 3752–3757.

33 Hamady, M., Walker, J.J., Harris, J.K. *et al.* (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. Methods*, **5** (3), 235–237.

34 Garst, A.D., Bassalo, M.C., Pines, G. *et al.* (2016) Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. *Nat. Biotechnol.*, **35**, 48–55.

35 Haliburton, J.R., Shao, W., Deutschbauer, A. *et al.* (2017) Genetic interaction mapping with microfluidic-based single cell sequencing. *PLoS One*, **12** (2), e0171302.

36 Zeitoun, R.I., Garst, A.D., Degen, G.D. *et al.* (2015) Multiplexed tracking of combinatorial genomic mutations in engineered cell populations. *Nat. Biotechnol.*, **33** (6), 631–637.

37 Zeitoun, R.I., Pines, G., Grau, W.C., and Gill, R.T. (2017) Quantitative tracking of combinatorially engineered populations with multiplexed binary assemblies. *ACS Synth. Biol.*, **6**, 619–627.

38 Covert, M.W., Schilling, C.H., Famili, I. *et al.* (2001) Metabolic modeling of microbial strains in silico. *Trends Biochem. Sci*, **26** (3), 179–186.

39 Tan, Y. and Liao, J.C. (2012) Metabolic ensemble modeling for strain engineers. *Biotechnol. J.*, **7** (3), 343–353.

40 Wang, H.H., Isaacs, F.J., Carr, P.A. *et al.* (2009) Programming cells by multiplex genome engineering and accelerated evolution. *Nature*, **460** (7257), 894–898.

41 Toscano, M.D., Woycechowsky, K.J., and Hilvert, D. (2007) Minimalist active-site redesign: teaching old enzymes new tricks. *Angew. Chem. Int. Ed.*, **46** (18), 3212–3236.

42 Jiang, W., Bikard, D., Cox, D. *et al.* (2013) RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.*, **31** (3), 233–239.

43 Oh, J.-H. and van Pijkeren, J.-P. (2014) CRISPR–Cas9-assisted recombineering in *Lactobacillus reuteri*. *Nucleic Acids Res.*, **42**, e131.

44 Pines, G., Pines, A., Garst, A.D. *et al.* (2015) Codon compression algorithms for saturation mutagenesis. *ACS Synth. Biol.*, **4** (5), 604–614.

45 Jinek, M., Chylinski, K., Fonfara, I. *et al.* (2012) A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337** (6096), 816–821.

3

Site-Directed Genome Modification with Engineered Zinc Finger Proteins

Lauren E. Woodard[1,2], Daniel L. Galvan[3], and Matthew H. Wilson[1,2]

[1] *Department of Veterans Affairs, Nashville, TN 37212, USA*
[2] *Vanderbilt University Medical Center, Department of Medicine, Department of Pharmacology, Nashville, TN 37232, USA*
[3] *University of Texas at MD Anderson Cancer Center, Section of Nephrology, Houston, TX 77030, USA*

The ability to precisely manipulate genomic DNA in living cells in a site-specific manner has revolutionized biomedical research. Site-specific genomic modification has greatly advanced preclinical research by creating invaluable cellular and animal models of disease and is currently in clinical trials for therapeutic application. Zinc finger proteins (ZFPs) represent a class of proteins that can be engineered to manipulate user-defined chromosomal DNA targets with a high degree of specificity. Zinc finger nucleases (ZFNs) cause double-stranded breaks (DSBs) at precise genomic locations that can induce deletions, insertions, translocations, and/or point mutations in the genomic DNA via endogenous DNA repair mechanisms. ZFPs fused to recombinases or transposases act in an autonomous manner without the need to induce toxic DSBs. This chapter represents an overview of ZFPs, the various methods available to researchers for engineering them, options for genomic modifications, methods for validation of genomic modifications, an overview of options for delivery to cells, and some novel ways that zinc fingers (ZFs) are being used for genomic alteration.

3.1 Introduction to Zinc Finger DNA-Binding Domains and Cellular Repair Mechanisms

3.1.1 Zinc Finger Proteins

The Cys_2-His_2 ZF domain makes up the most common DNA-binding domain structure in eukaryotes [1]. Structural determination of ZF domains bound to DNA has enabled rational design of proteins to bind targeted DNA sequences [1]. Such engineered ZFP domains can be fused to other protein domains with differing capabilities to create enzymes capable of targeted cleavage of DNA and other targeted genomic effects [2]. ZFPs can be engineered with a high degree of specificity for unique genomic elements [3]. This chapter mainly focuses on the

Synthetic Biology: Parts, Devices and Applications, First Edition. Edited by Christina Smolke.

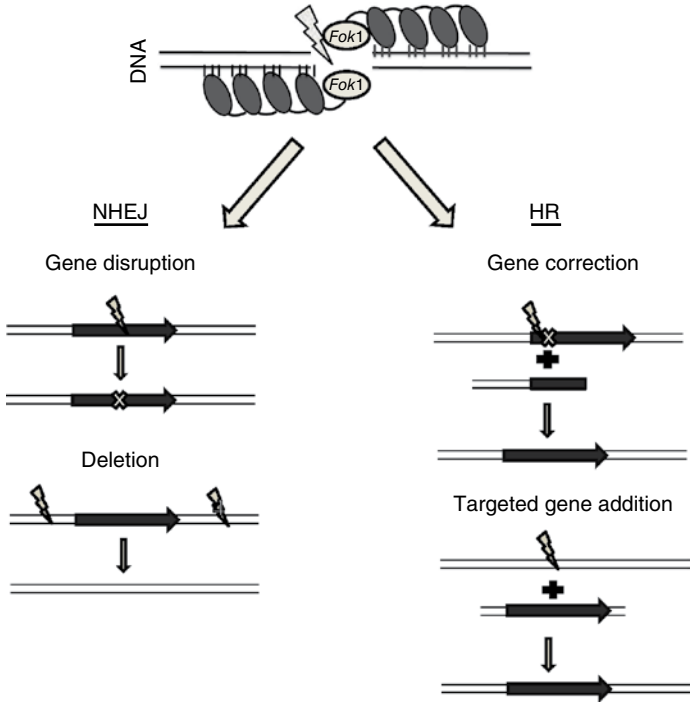


Figure 3.1 Targeted genomic modification using zinc finger nucleases (ZFNs). A pair of ZFPs fused to the *Fok*I nuclease domain is designed to target opposite strands of DNA. When dimerization of the *Fok*I domain occurs following ZF binding, a double-strand break (DSB) is created in the DNA (shown by lightning bolt). The cell chooses to use either the nonhomologous end joining (NHEJ) or homologous recombination (HR) pathway to repair the DSB. NHEJ can be used for gene disruption via targeted mutagenesis using one pair of ZFNs or deletions/inversions with two pairs of ZFNs. If gene correction via HR is desired, a homologous template sequence is provided that will be used by the cellular HR machinery replace the endogenous sequence near the DSB. Alternatively, targeted gene addition at or near the site of the targeted DNA cleavage can be achieved by flanking the sequence to be inserted with homologous arms.

use of engineered ZFPs called zinc finger nucleases for site-directed genomic modification through targeted DNA cleavage. When a targeted double-stranded DNA break is engineered, endogenous repair subsumes via either homologous recombination (HR) or nonhomologous end joining (NHEJ) [3] (Figure 3.1).

3.1.2 Homologous Recombination

HR is a process of exchanging shared DNA sequences between sister chromatids. HR naturally occurs in a diverse range of organisms, from bacteria to humans. The HR process has two major purposes: (i) the protection of somatic genomes through DSB repair to prevent mutations that could result in cell death or cancer and (ii) to increase the genetic diversity of the next generation through recombination of the parental chromosomes in each gamete during meiosis. HR has been

proven to be very useful for the genomic manipulation of yeast, where rates of HR are naturally high [4]. In contrast, mammalian cells provided with a homologous DNA template have extremely low background rates of HR: only one in a million somatic cells shows evidence of HR-mediated repair following homologous DNA introduction [5, 6]. To stimulate repair, DSBs may be introduced to encourage the cell to repair the DNA using the abundantly available homologous DNA template [7]. ZFN-induced DSBs stimulate HR at target sequences. In this way ZFNs may be used to introduce or correct point mutations in a seamless manner without additional sequences, making this strategy preferred to gene addition strategies when possible. Unfortunately, the frequency of HR is cell-type dependent and cell division is required, ruling out many potential applications [8, 9]. Despite the limitations, this method has been applied for genomic manipulation of many species across the kingdoms of life, including bacteria, yeast, plants, and mice, as well as in human cells for seamless gene correction [10].

3.1.3 Non-homologous End Joining

NHEJ is a system of DSB repair that acts by directly ligating the ends of linear DNA. If the break is staggered and homologous sequences exist on either side, an accurate repair can be made by sensing the homology [11]. However, if there are no homologous strands, NHEJ can still mediate repair to stitch back together the DNA via the protein Ku70 [10, 12]. In this case, there is usually the gain or loss of a few base pairs of DNA resulting from the chemical repair of the free ends of the DNA. Under highly stressful conditions such as ultraviolet radiation, toxins, radioactivity, or desiccation, the cell could suffer multiple DSBs. In this case, when there are more than two free linear ends, the cell may ligate the incorrect ends together, resulting in chromosomal rearrangements or large deletions. Such major insults can result in a cancer-causing phenotype or more likely cell death. Even with this possibility, NHEJ is still highly conserved due to the huge advantage to the organism of having a DNA repair mechanism that does not rely on the presence of homologous sequences on the sister chromatid.

Gene deletion or addition can be achieved with NHEJ, while the elegant seamless gene correction or addition strategies require HR. The majority of the cell types comprising an adult human, which are the cells that are most desirable to be targeted for correction in a gene therapy setting, prefer NHEJ over HR. In transfected cells, using I-SceI to induce a DSB and the sister chromatid for HR rarely results in gene correction, making NHEJ the easier goal to achieve. Demonstrating this point, a clinical trial has been completed using a ZFN to knock out the CCR5 receptor to block HIV infection [13, 14], while there are no clinical trials underway that rely upon HR-mediated gene correction.

Either of these repair strategies can be exploited using ZFNs for targeted genomic modification. NHEJ can be used for targeted mutagenesis of chromosomal elements [15]. HR can be used for targeted DNA repair or gene addition by providing a template strand of DNA homologous to the targeted site of DNA cleavage [16–18]. Therefore, targeted DNA cleavage can be used to achieve targeted mutagenesis, targeted DNA repair, or targeted DNA addition at specific genomic sites.

3.2 Approaches for Engineering or Acquiring Zinc Finger Proteins

The most common approach for targeted genomic DNA cleavage via ZFPs is to use ZFNs [19–21]. Simplistically, a ZFN involves fusion of a ZFP to a nuclease domain via a short flexible linker sequence. The simplest ZFN combines a naturally occurring ZFP with the linker and FokI nuclease domains to target its native binding site [17]. However, ZFNs can also be rationally designed to target a wide range of sequences for a greater number of applications.

ZF motifs are 30-amino-acid protein domains that chelate a zinc ion. They bind to DNA by insertion of an alpha helix into the major groove of the DNA to probe the DNA sequence [22]. Naturally occurring ZFs may be mutated to alter binding specificity [23]. The DNA sequence that will be bound is defined by certain amino acids [24, 25]. These ZFs can be combined into strings of 3, 4, 5, or 6 ZFs to bind increasingly long DNA sequences to enhance the specificity of the interaction [26–28]. The availability of motifs that recognize triplet sequences is a limiting factor in ZF design, as the ZFN pair should be designed such that they will create a DSB as close to the site of desired HR as possible [29].

Most restriction enzymes cleave palindromic sequences through coupled DNA-binding and cleavage events. The *Fok*I endonuclease is different in that it cuts DNA between two binding sites that can be 9–18 bp in length on opposite DNA strands. *Fok*I contains two separate domains: the N-terminal domain is involved in sequence recognition, while the C-terminal domain contains a nuclease [30, 31]. *Fok*I is unique in that single amino acid substitutions resulted in the decoupling of sequence recognition and cleavage [32, 33], allowing the nuclease domain to be isolated and fused to other DNA-binding domains. In addition, dimerization of the nuclease domain is required for DNA cleavage to occur [34]. The ZFN architecture has been improved such that cleavage by the enzyme requires a heterodimer to be formed, preventing the off-target events that could result from homodimer formation [35]. A recent study reported a multi-reporter selection system to identify ZFNs with high degrees of activity at the desired site and negligible activity at similar off-target sites in the genome [36]. Refinements through mutagenesis and DNA shuffling have made the *Fok*I cleavage domain 15-fold more active and 6-fold more specific [37]. Therefore, ZFN pairs can be engineered such that DNA binding by each ZFN mediates *Fok*I nuclease dimerization between ZFP binding sites, resulting in targeted DNA cleavage [38, 39] (Figure 3.1).

A potential design limitation when designing a ZFP is the lack of availability of ZF motifs to recognize every triplet sequence [29]. In order to make longer strings of 6 ZFs, the longer recognition site (18–19 bp) must have ZFs that can recognize the entire sequence. There are several options available to investigators for engineering ZFPs for this purpose, and these include modular assembly, a selection method termed "OPEN," context-dependent assembly termed "CoDA," and a proprietary system available from Sigma-Aldrich. These differing approaches are discussed in brief later.

3.2.1 Modular Assembly

Modular assembly for engineering ZFPs utilizes a combination of validated ZF modules that each targets a separate DNA triplet. This combination allows for longer DNA sequences with higher probabilities of being unique in the genome to be targeted [22]. These modules can be combined by drawing from toolkits available from Barbas [40], ToolGen [41], and Sigma-Aldrich. Modules are then strung together for *in silico* predicted targeted binding of the desired DNA sequence [41, 42]. These end-effect ZFP sequences can be retrieved from a web server that utilizes known ZF binding to DNA triplets and designs the engineered ZFP *in silico.* The modules can be tied together using molecular biology techniques or gene synthesis. A potential drawback of this method is that the specificity of each ZF module can depend on both the context of the surrounding DNA target sequence and the other protein components that it is linked to [43]. Along these lines, modular assembly-produced four-finger ZFs outperform three-finger ZFs [41]. For these reasons, *in silico* prediction alone is not ideal for most applications. Modular assembly should be combined with a selection method to test many predicted ZFPs to find the one with the most desirable features for expression and binding to the desired target sequence, such as high activity and specificity.

3.2.2 OPEN and CoDA Selection Systems

Several selection methods have been devised to address the limitations of *in silico* modular assembly by relying on screening for optimal binding capabilities from large libraries of potential ZFPs. Initially, partially randomized ZF arrays were screened in large pools by phage display to select for those that could effectively bind to the desired DNA sequence [44, 45]. Pabo's group devised a successful strategy to gradually extend the ZFP by adding and optimizing each finger individually [46]. More recently, oligomerized pool engineering, or "OPEN" [47], derives ZFPs from randomized libraries. Each finger in a three-finger ZFP was randomized and the resulting library was screened using low-stringency selection methods [18]. The resultant clones were then picked to generate a pool of potential ZFPs that was further recombined by swapping the fingers [18]. These randomized, then recombined, three-finger ZFPs were selected for the optimal combination of fingers to bind to the desired target site [18]. While OPEN is available to all researchers, screening the large libraries that are generated requires a serious time investment and some skilled knowledge of the components involved. This has limited the adoption of OPEN. The latest generation of ZFN assembly is termed context-dependent assembly [48], which takes into account interactions between ZFPs while using modular assembly [19]. The CoDA approach can be used to create an array of viable ZFP options for many target sites with a similar efficiency to OPEN but is easier and faster to use [19]. CoDA relies upon arrays of previously validated three-finger ZFPs that share a common middle finger and are shuffled via this homologous sequence to create an array [19]. All of the software and reagents required to implement CoDA are publicly available. The Zinc Finger Consortium offers web-based

tools for evaluating for ZFN target sites within a genomic DNA region for both OPEN and CoDA at the www.zincfingers.org website. Because OPEN and CoDA rely upon previously validated ZFPs, there are some sequences that cannot be targeted using these methods. Based on the failure rates and time investment for each approach, investigators should first consider CoDA, then OPEN, and then modular assembly only if the target sequence is unavailable through CoDA or OPEN. Supporting this recommendation, recent computational studies have suggested that binding of the ZFP to the DNA sequence is better thought of as synergistic rather than strictly modular [49, 50].

3.2.3 Purchase via Commercial Avenues

Engineered ZFPs are also available commercially. Sangamo Therapeutics, Inc. developed a proprietary archive of engineered ZFs early on but has not made this information public, although they have published some of the details regarding their ZFN engineering platform [48]. Currently, the simplest means of obtaining a ZFN pair to a novel target sequence is by purchasing a custom protein. Sangamo licensed its proprietary methodology to Sigma-Aldrich, which has marketed the technology as the CompoZr Zinc Finger Nuclease platform. Pre-validated ZFNs to the rat and mouse Rosa26 locus as well as the human AAVS1 safe harbor site present the most cost-effective option. These would allow the investigator to place transgenes at known genomic locations that will not interfere with genomic function, are commonly used, and are known "safe harbor" sites. Additionally, ZFNs to target an abundance of specific human, mouse, and rat genes are available at a more reasonable cost as compared with custom target options; the complete list of the genes is available online at www.sigmaaldrich.com. Custom ZFNs designed to target novel sequences require increased time and are produced at a much greater cost. A major advantage of using a commercial service to design a custom ZFN is the timeframe of delivery in less than 3 months. For most research investigators, use of the clustered regularly interspaced short palindromic repeat (CRISPR)/Cas9 system of targeted integration is now the fastest and most cost-effective method by which to initiate a nuclease-driven project [10]. Later on, purchased or designed ZFNs may be integrated into the molecular toolkit for intellectual property, reproducibility, or other experimental reasons.

3.3 Genome Modification with Zinc Finger Nucleases

Engineered ZFNs can be used for a variety of genome alterations. These can be categorized as dependent on either HR or NHEJ. HR-based alterations include targeted addition of DNA sequence to the genome [16–18] through introduction of a new sequence flanked by homologous arms or targeted base-pair changes achieved by supplying a homologous template with the desired alteration. NHEJ-based changes do not require the introduction of homologous sequences and include gene disruption strategies that take advantage of the infidelity of NHEJ repair mechanisms. Another NHEJ strategy involves supplying ZFN pairs for

two sites to introduce two DSBs that may result in a large deletion or chromosomal translocation [51].

HR-based methods can be used to introduce transgene sequences or small mutations at target sites by providing a homologous donor DNA template. This template should include 700 bp homology arms if it is typical double-stranded circular DNA [27]. For linear DNA, only 50 bp of homology is required [19]. Single-stranded DNA oligonucleotides have also been used to achieve point mutagenesis, deletions, or insertions [52, 53]. As compared with HR methods that involve simply introducing the homologous sequence with the desired mutation, introducing a targeted double-stranded DNA break enhances the efficiency of genome editing by many orders of magnitude [18, 54–57]. HR-based methods do not work in every cell type, however, as they require the presence of the HR machinery, which is only available during the S and G2 phases of the cell cycle just prior to mitosis. Strategies employing HR can be achieved at desired rates in early stem cells. However, HR-directed genomic modification cannot be achieved at appreciable rates in many differentiated cell types because these cells are not dividing and do not have the HR machinery available. This presents a major roadblock for the design of a gene therapy-type strategy based on ZFN-induced HR. There are also species-specific differences in the frequency of HR to consider: for example, mouse embryonic stem (ES) cells are more prone to HR and thus easier to modify than human ES cells [58, 59].

HR is an elegant and seamless method to create perfectly tailored DNA sequences in the genome, but many somatic cells rely on the NHEJ repair pathway instead. NHEJ-based gene disruption is much easier to achieve than HR, although the resulting mutations are not predictable. Thus far the only clinical trial to date using a genome modification system based on nucleases uses a ZFN pair to disrupt the *CCR5* locus [13, 14]. Since the CCR5 cell surface receptor is required for most HIV infection, disruption of this locus was used to create CD4 T cells that are unable to be infected by the HIV [13, 14]. The phase I clinical trial indicated that patients infused with T cells that were modified via ZFN technology to lack functional CCR5 receptors exhibited a slower rate of decline in the modified CD4 T cells relative to unmodified T cells [13, 14]. Among the 12 clinical trial participants, one serious adverse event was reported of a patient suffering fever, chills, and joint pain the day following infusion [14]. Nevertheless, the authors concluded that the autologous CD4 T-cell infusions were safe [14]. They also found that the blood level of HIV DNA decreased in most patients and one out of four patients tested had no detectable traces of HIV RNA in the tested samples, suggesting efficacy [14].

In addition to gene disruption via NHEJ, pairs of ZFNs may be designed to induce chromosomal translocations [60] by causing two simultaneous DSBs at desired locations. This technique could be used to study translocations that are important for cancer formation. The same two-DSB strategy will also produce deletions of up to 15 Mb [51]. These deletions could be used to remove an exon, an entire genomic locus, or even a number of genes from the genome. Therefore, multiple options exist both for achieving the desired genome modification and the methods by which to achieve those modifications using ZFNs.

3.4 Validating Zinc Finger Nuclease-Induced Genome Alteration and Specificity

Methods have been developed for monitoring for endogenous gene modification. One such assay evaluates for DNA DSB repair via using the Surveyor nuclease [35]. This assay involves three steps: (i) polymerase chain reaction (PCR) amplification of the region of interest and annealing of the strands to form homoduplexes (no mismatches) and heteroduplexes (containing mismatches), (ii) cleavage of the mismatched heteroduplexes by the Surveyor nuclease, and (iii) fragment size evaluation to determine if mismatched DNA was present [61]. By only cleaving annealed complexes containing both the mutated and wild-type DNA after amplification, the Surveyor nuclease can used to estimate the level of mutagenesis mediated by the ZFNs (Figure 3.2).

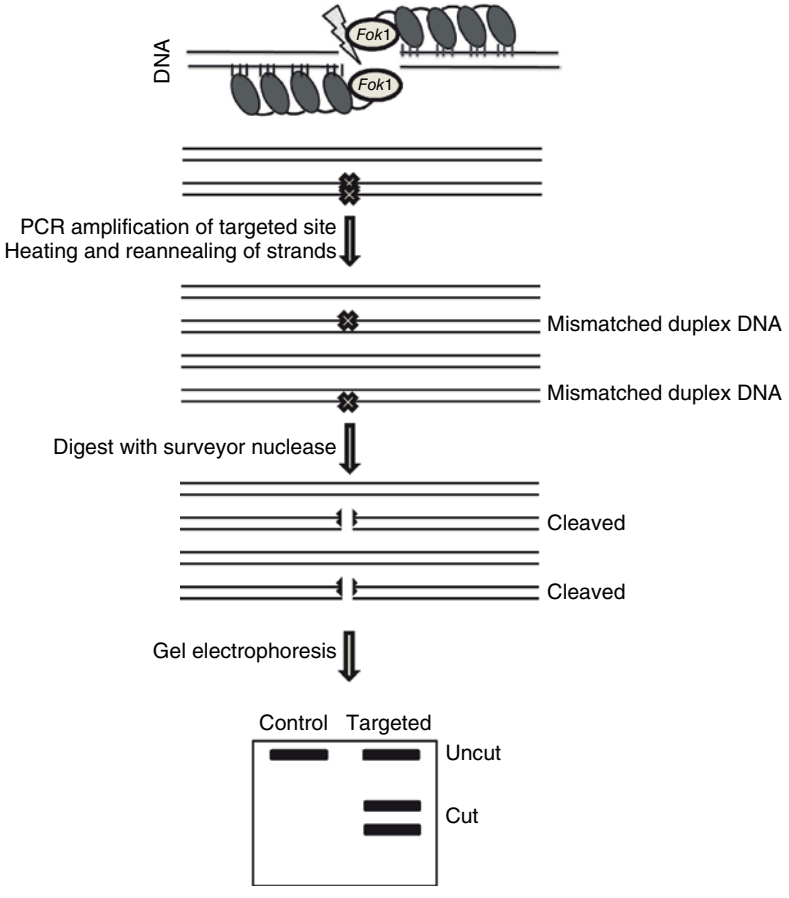


Figure 3.2 A nuclease assay for detecting gene targeting. ZFNs are used to create a targeted DNA DSB. PCR is used to amplify the targeted sites and the DNA is heated and cooled to reanneal the strands, creating mismatched heteroduplexes of DNA. The Surveyor nuclease cleaves the heteroduplexes only at the sites of mismatched DNA, leaving the homodimers unmodified. Gel electrophoresis can then be used to observe and quantitate the efficiency of gene targeting using ZFNs at the genomic level.

The Surveyor nuclease can be used to evaluate if the desired modification was achieved, but it cannot be used to evaluate the number of off-target events in an unbiased manner. The specificity of genome modification is also highly important, whether one is selecting a gene knockout by limiting dilution or determining the percentage of off-target modifications in a pool to make inferences about gene therapy safety. To probe the specificity of the ZFP for the target DNA-binding site, *in vitro* binding profiles are experimentally derived [62]. Then, *in silico* prediction can be used to determine a number of predicted off-target sites to be PCR amplified. The Surveyor assay may be used to evaluate off-target cleavage at the predicted loci. Because this technique is limited by the successful prediction of the off-target binding sites, large-scale sequencing methods may be preferable as they provide a more comprehensive view of all off-target events in the genomic DNA [63, 64]. Whole-exome sequencing and next-generation sequencing methods have become widely available and more commonplace in recent years [65]. These sequencing methods permit unbiased whole-genome analysis of ZFN specificity. However, small numbers of off-target events may not be effectively found by this or any method, so practical assessment of the transformation and clonal expansion of treated cells may be performed by established methods such as a soft agar assay are also advisable for development of ZFN-based clinical products [66].

3.5 Methods for Delivering Engineered Zinc Finger Nucleases into Cells

The ability to perform targeted genome modification using ZFNs is dependent on the delivery of the ZFNs into target cells and into the nucleus. The ZFN genes may be introduced into the cell by viral or nonviral methods. Nonviral methods to transfect cells *ex vivo* include lipophilic reagents and electroporation. Electroporation of plasmid DNA or RNA has proven effective, though electroporation can be toxic to cells and is less efficient than other methods such as viral delivery [67]. Viral delivery has been successful, including adenovirus [59], integrase-defective lentivirus [68], and adeno-associated virus [69–71]. Viral delivery can be used for both the delivery of the ZFN and homologous DNA if HR-directed modifications are desired. More recently, ZFN protein has been shown to be capable of traversing cell membranes to achieve genome editing [72]. Ultimately, the delivery methodology use for ZFN-mediated genome modification will depend on the cell type targeted and whether or not the cells will be modified *in vitro* or *in vivo*.

3.6 Zinc Finger Fusions to Transposases and Recombinases

ZFNs comprise the more characterized class of proteins for site-directed genome modification. However, all nucleases, including ZFNs, have serious limitations. Difficulties in measuring the rates of off-target DNA cleavage, the dependence on cellular DNA repair machinery, high levels of DSB-induced toxicity leading to

cell death, and the requirement for cell division are just some of the problems caused by inducing free DSBs in the cell. DSBs are associated with carcinogenic agents and can cause undesired chromosomal translocations [73], although there have not been any reports to date of ZFNs causing cancer. Enzymes such as recombinases and transposases are capable of DNA excision and integration autonomously, without free DSBs and their negative aspects. However, transposases require very short sequences for integration, usually 2–8 bp, making their integration essentially random [74]. Fusion of ZF DNA-binding domains to recombinases [75–77] and transposases [78, 79] has resulted in successful redirection of the integration events to varying degrees. Recombinases have built-in DNA specificity and thus require reengineering to target user-defined chromosomal targets [75–77]. Transposase fusions do not require such engineering since their target sites are so short. ZF–transposase fusions are sometimes highly active [78, 80, 81]. However, these systems require further refinement. Firstly, transposase fusions have not yet demonstrated a high level of specificity in genomic targeting because the transposase portion of the ZFP transposes in a manner that is independent of the ZF DNA-binding domain. In order to increase specificity, one idea is to mutate the ZF–transposase fusions such that the transposase domain is kept inactive until the ZF portion binds the DNA. Secondly, transposase ZFPs require the presence of their short transposase target site in close proximity to the site recognized by the ZF [79], placing a limit on the available target sites. Finally, despite many advances, effective engineering of a ZFP to target a unique genomic locus has not yet been accomplished. Attempts to target the checkpoint kinase-2 (CHK2), the ROSA26 locus, and the L-gulono-γ-lactone oxidase pseudogene (GULOP) were unable to produce successful targeting in cells [79, 82]. Further development of proteins other than ZFNs for genomic targeting should lead to diverse technologies capable of site-specific gene addition, even in cells not actively dividing.

3.7 Conclusions

ZFNs are a proven tool for targeting endogenous loci in the genome, while ZFPs have the potential for user-defined modification of chromosomal targets without DSBs. Over the years "open" access to ZFP engineering tools together with commercial availability led to more widespread use. However, while ZFNs and other nucleases began the field of targeted genome modification, one might expect for the focus on ZFNs to decrease in the coming years as the ease and simplicity of working with the CRISPR/Cas9 system displaces the older, more expensive, and time-consuming ZFN platform. The Cas9 system can attribute the exponential pace of its development to the established systems and assays that were developed for engineering and testing ZFNs. As clinical trials usually take over a decade to reach the clinic and the ZFNs have a different set of patents governing their use, it is still quite possible that ZFN-based drugs could become approved for therapeutic use at some point in the near future. ZFPs will continue to be important tools for genome engineering to ask critical biological questions as well as development of novel therapeutics to improve human health. Time, and

much research, will tell which genome engineering platform will be most fruitful for desired applications or therapeutic goals.

References

1 Pabo, C.O., Peisach, E., and Grant, R.A. (2001) Design and selection of novel Cys_2His_2 zinc finger proteins. *Annu. Rev. Biochem.*, **70**, 313–340.
2 Gersbach, C.A., Gaj, T., and Barbas, C.F. 3rd (2014) Synthetic zinc finger proteins: the advent of targeted gene regulation and genome modification technologies. *Acc. Chem. Res.*, **47**, 2309–2318.
3 Wyman, C. and Kanaar, R. (2006) DNA double-strand break repair: all's well that ends well. *Annu. Rev. Genet.*, **40**, 363–383.
4 Hinnen, A., Hicks, J.B., and Fink, G.R. (1978) Transformation of yeast. *ProcProc. Natl. Acad. Sci. U.S.A.*, **75**, 1929–1933.
5 Moynahan, M.E. and Jasin, M. (1997) Loss of heterozygosity induced by a chromosomal double-strand break. *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 8988–8993.
6 Richardson, C., Moynahan, M.E., and Jasin, M. (1998) Double-strand break repair by interchromosomal recombination: suppression of chromosomal translocations. *Genes Dev.*, **12**, 3831–3842.
7 Smithies, O., Gregg, R.G., Boggs, S.S., Koralewski, M.A. *et al.* (1985) Insertion of DNA sequences into the human chromosomal β-globin locus by homologous recombination. *Nature*, **317**, 230–234.
8 Resnick, M.A. and Moore, P.D. (1979) Molecular recombination and the repair of DNA double-strand breaks in CHO cells. *Nucleic Acids Res.*, **6**, 3145–3160.
9 Rahman, S.H., Bobis-Wozowicz, S., Chatterjee, D., Gellhaus, K. *et al.* (2013) The nontoxic cell cycle modulator indirubin augments transduction of adeno-associated viral vectors and zinc-finger nuclease-mediated gene targeting. *Hum. Gene Ther.*, **24**, 67–77.
10 Kim, H. and Kim, J.S. (2014) A guide to genome engineering with programmable nucleases. *Nat. Rev. Genet.*, **15**, 321–334.
11 Moore, J.K. and Haber, J.E. (1996) Capture of retrotransposon DNA at the sites of chromosomal double-strand breaks. *Nature*, **383**, 644–646.
12 Boulton, S.J. and Jackson, S.P. (1996) Saccharomyces cerevisiae Ku70 potentiates illegitimate DNA double-strand break repair and serves as a barrier to error-prone DNA repair pathways. *EMBO J.*, **15**, 5093–5103.
13 Maier, D.A., Brennan, A.L., Jiang, S., Binder-Scholl, G.K. *et al.* (2013) Efficient clinical scale gene modification via zinc finger nuclease-targeted disruption of the HIV co-receptor CCR5. *Hum. Gene Ther.*, **24**, 245–258.
14 Tebas, P., Stein, D., Tang, W.W., Frank, I. *et al.* (2014) Gene editing of CCR5 in autologous CD4 T cells of persons infected with HIV. *N. Engl. J. Med.*, **370**, 901–910.
15 Bibikova, M., Golic, M., Golic, K.G., and Carroll, D. (2002) Targeted chromosomal cleavage and mutagenesis in *Drosophila* using zinc-finger nucleases. *Genetics*, **161**, 1169–1175.
16 Bibikova, M., Beumer, K., Trautman, J.K., and Carroll, D. (2003) Enhancing gene targeting with designed zinc finger nucleases. *Science*, **300**, 764.

17 Bibikova, M., Carroll, D., Segal, D.J., Trautman, J.K. *et al.* (2001) Stimulation of homologous recombination through targeted cleavage by chimeric nucleases. *Mol. Cell. Biol.*, **21**, 289–297.

18 Porteus, M.H. and Baltimore, D. (2003) Chimeric nucleases stimulate gene targeting in human cells. *Science*, **300**, 763.

19 Orlando, S.J., Santiago, Y., DeKelver, R.C., Freyvert, Y. *et al.* (2010) Zinc-finger nuclease-driven targeted integration into mammalian genomes using donors with limited chromosomal homology. *Nucleic Acids Res.*, **38**, e152.

20 Carroll, D. (2011) Genome engineering with zinc-finger nucleases. *Genetics*, **188**, 773–782.

21 Rahman, S.H., Maeder, M.L., Joung, J.K., and Cathomen, T. (2011) Zinc-finger nucleases for somatic gene therapy: the next frontier. *Hum. Gene Ther.*, **22**, 925–933.

22 Pavletich, N.P. and Pabo, C.O. (1991) Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science*, **252**, 809–817.

23 Jamieson, A.C., Kim, S.H., and Wells, J.A. (1994) In vitro selection of zinc fingers with altered DNA-binding specificity. *Biochemistry*, **33**, 5689–5695.

24 Shi, Y. and Berg, J.M. (1995) A direct comparison of the properties of natural and designed zinc-finger proteins. *Chem. Biol.*, **2**, 83–89.

25 Elrod-Erickson, M. and Pabo, C.O. (1999) Binding studies with mutants of Zif268. Contribution of individual side chains to binding affinity and specificity in the Zif268 zinc finger-DNA complex. *J. Biol. Chem.*, **274**, 19281–19285.

26 Beerli, R.R., Segal, D.J., Dreier, B., and Barbas, C.F. III (1998) Toward controlling gene expression at will: specific regulation of the *erbB-2/HER-2* promoter by using polydactyl zinc finger proteins constructed from modular building blocks. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 14628–14633.

27 Liu, Q., Segal, D.J., Ghiara, J.B., and Barbas, C.F. III (1997) Design of polydactyl zinc-finger proteins for unique addressing within complex genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 5525–5530.

28 Kim, J.S. and Pabo, C.O. (1998) Getting a handhold on DNA: design of poly-zinc finger proteins with femtomolar dissociation constants. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 2812–2817.

29 Durai, S., Mani, M., Kandavelou, K., Wu, J. *et al.* (2005) Zinc finger nucleases: custom-designed molecular scissors for genome engineering of plant and mammalian cells. *Nucleic Acids Res.*, **33**, 5978–5990.

30 Li, L., Wu, L.P., and Chandrasegaran, S. (1992) Functional domains in *Fok*I restriction endonuclease. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 4275–4279.

31 Li, L., Wu, L.P., Clarke, R., and Chandrasegaran, S. (1993) C-terminal deletion mutants of the *Fok*I restriction endonuclease. *Gene*, **133**, 79–84.

32 Wah, D.A., Bitinaite, J., Schildkraut, I., and Aggarwal, A.K. (1998) Structure of *Fok*I has implications for DNA cleavage. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 10564–10569.

33 Wah, D.A., Hirsch, J.A., Dorner, L.F., Schildkraut, I. *et al.* (1997) Structure of the multimodular endonuclease *Fok*I bound to DNA. *Nature*, **388**, 97–100.

34 Bitinaite, J., Wah, D.A., Aggarwal, A.K., and Schildkraut, I. (1998) *Fok*I dimerization is required for DNA cleavage. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 10570–10575.

35 Miller, J.C., Holmes, M.C., Wang, J., Guschin, D.Y. *et al.* (2007) An improved zinc-finger nuclease architecture for highly specific genome editing. *Nat. Biotechnol.*, **25**, 778–785.

36 Oakes, B.L., Xia, D.F., Rowland, E.F., Xu, D.J. *et al.* (2016) Multi-reporter selection for the design of active and more specific zinc-finger nucleases for genome editing. *Nat. Commun.*, **7**, 10194.

37 Guo, J., Gaj, T., and Barbas, C.F. III (2010) Directed evolution of an enhanced and highly efficient *Fok*I cleavage domain for zinc finger nucleases. *J. Mol. Biol.*, **400**, 96–107.

38 Mani, M., Smith, J., Kandavelou, K., Berg, J.M. *et al.* (2005) Binding of two zinc finger nuclease monomers to two specific sites is required for effective double-strand DNA cleavage. *Biochem. Biophys. Res. Commun.*, **334**, 1191–1197.

39 Smith, J., Bibikova, M., Whitby, F.G., Reddy, A.R. *et al.* (2000) Requirements for double-strand cleavage by chimeric restriction enzymes with zinc finger DNA-recognition domains. *Nucleic Acids Res.*, **28**, 3361–3369.

40 Gonzalez, B., Schwimmer, L.J., Fuller, R.P., Ye, Y. *et al.* (2010) Modular system for the construction of zinc-finger libraries and proteins. *Nat. Protoc.*, **5**, 791–810.

41 Kim, H.J., Lee, H.J., Kim, H., Cho, S.W. *et al.* (2009) Targeted genome editing in human cells with zinc finger nucleases constructed via modular assembly. *Genome Res.*, **19**, 1279–1288.

42 Mandell, J.G. and Barbas, C.F. III (2006) Zinc finger tools: custom DNA-binding domains for transcription factors and nucleases. *Nucleic Acids Res.*, **34**, W516–W523.

43 Ramirez, C.L., Foley, J.E., Wright, D.A., Muller-Lerch, F. *et al.* (2008) Unexpected failure rates for modular assembly of engineered zinc fingers. *Nat. Methods*, **5**, 374–375.

44 Choo, Y. and Klug, A. (1994) Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 11168–11172.

45 Choo, Y. and Klug, A. (1994) Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 11163–11167.

46 Greisman, H.A. and Pabo, C.O. (1997) A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science*, **275**, 657–661.

47 Maeder, M.L., Thibodeau-Beganny, S., Osiak, A., Wright, D.A. *et al.* (2008) Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Mol. Cell*, **31**, 294–301.

48 Doyon, Y., McCammon, J.M., Miller, J.C., Faraji, F. *et al.* (2008) Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. *Nat. Biotechnol.*, **26**, 702–708.

49 Dutta, S., Madan, S., Parikh, H., and Sundar, D. (2016) An ensemble micro neural network approach for elucidating interactions between zinc finger proteins and their target DNA. *BMC Genomics*, **17**, 1033.

50 Dutta, S., Madan, S., and Sundar, D. (2016) Exploiting the recognition code for elucidating the mechanism of zinc finger protein-DNA interactions. *BMC Genomics*, **17**, 1037.

51 Lee, H.J., Kim, E., and Kim, J.S. (2010) Targeted chromosomal deletions in human cells using zinc finger nucleases. *Genome Res.*, **20**, 81–89.

52 Chen, F., Pruett-Miller, S.M., Huang, Y., Gjoka, M. *et al.* (2011) High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. *Nat. Methods*, **8**, 753–755.

53 Radecke, S., Radecke, F., Cathomen, T., and Schwarz, K. (2010) Zinc-finger nuclease-induced gene repair with oligodeoxynucleotides: wanted and unwanted target locus modifications. *Mol. Ther.*, **18**, 743–753.

54 Choulika, A., Perrin, A., Dujon, B., and Nicolas, J.F. (1995) Induction of homologous recombination in mammalian chromosomes by using the I-*Sce*I system of *Saccharomyces cerevisiae*. *Mol. Cell Biol.*, **15**, 1968–1973.

55 Brenneman, M., Gimble, F.S., and Wilson, J.H. (1996) Stimulation of intrachromosomal homologous recombination in human cells by electroporation with site-specific endonucleases. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 3608–3612.

56 Donoho, G., Jasin, M., and Berg, P. (1998) Analysis of gene targeting and intrachromosomal homologous recombination stimulated by genomic double-strand breaks in mouse embryonic stem cells. *Mol. Cell Biol.*, **18**, 4070–4078.

57 Rouet, P., Smih, F., and Jasin, M. (1994) Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Mol. Cell Biol.*, **14**, 8096–8106.

58 Doetschman, T., Gregg, R.G., Maeda, N., Hooper, M.L. *et al.* (1987) Targetted correction of a mutant HPRT gene in mouse embryonic stem cells. *Nature*, **330**, 576–578.

59 Thomas, K.R. and Capecchi, M.R. (1987) Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells. *Cell*, **51**, 503–512.

60 Brunet, E., Simsek, D., Tomishima, M., DeKelver, R. *et al.* (2009) Chromosomal translocations induced at specified loci in human stem cells. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 10620–10625.

61 Guschin, D.Y., Waite, A.J., Katibah, G.E., Miller, J.C. *et al.* (2010) A rapid and general assay for monitoring endogenous gene modification. *Methods Mol. Biol.*, **649**, 247–256.

62 Perez, E.E., Wang, J., Miller, J.C., Jouvenot, Y. *et al.* (2008) Establishment of HIV-1 resistance in $CD4^+$ T cells by genome editing using zinc-finger nucleases. *Nat. Biotechnol.*, **26**, 808–816.

63 Gabriel, R., Lombardo, A., Arens, A., Miller, J.C. *et al.* (2011) An unbiased genome-wide analysis of zinc-finger nuclease specificity. *Nat. Biotechnol.*, **29**, 816–823.

64 Pattanayak, V., Ramirez, C.L., Joung, J.K., and Liu, D.R. (2011) Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. *Nat. Methods*, **8**, 765–770.

65 Yusa, K., Rashid, S.T., Strick-Marchand, H., Varela, I. *et al.* (2011) Targeted gene correction of α_1-antitrypsin deficiency in induced pluripotent stem cells. *Nature*, **478**, 391–394.

66 Hendel, A., Fine, E.J., Bao, G., and Porteus, M.H. (2015) Quantifying on- and off-target genome editing. *Trends Biotechnol.*, **33**, 132–140.

67 Torikai, H., Reik, A., Soldner, F., Warren, E.H. *et al.* (2013) Toward eliminating HLA class I expression to generate universal cells from allogeneic donors. *Blood*, **122**, 1341–1349.

68 Lombardo, A., Genovese, P., Beausejour, C.M., Colleoni, S. *et al.* (2007) Gene editing in human stem cells using zinc finger nucleases and integrase-defective lentiviral vector delivery. *Nat. Biotechnol.*, **25**, 1298–1306.

69 Handel, E.M., Gellhaus, K., Khan, K., Bednarski, C. *et al.* (2012) Versatile and efficient genome editing in human cells by combining zinc-finger nucleases with adeno-associated viral vectors. *Hum. Gene Ther.*, **23**, 321–329.

70 Li, H., Haurigot, V., Doyon, Y., Li, T. *et al.* (2011) *In vivo* genome editing restores haemostasis in a mouse model of haemophilia. *Nature*, **475**, 217–221.

71 Ellis, B.L., Hirsch, M.L., Porter, S.N., Samulski, R.J. *et al.* (2013) Zinc-finger nuclease-mediated gene correction using single AAV vector transduction and enhancement by Food and Drug Administration-approved drugs. *Gene Ther.*, **20**, 35–42.

72 Gaj, T., Guo, J., Kato, Y., Sirk, S.J. *et al.* (2012) Targeted gene knockout by direct delivery of zinc-finger nuclease proteins. *Nat. Methods*, **9**, 805–807.

73 Richardson, C. and Jasin, M. (2000) Frequent chromosomal translocations induced by DNA double-strand breaks. *Nature*, **405**, 697–700.

74 Woodard, L.E., Li, X., Malani, N., Kaja, A. *et al.* (2012) Comparative analysis of the recently discovered *hAT* transposon *TcBuster* in human cells. *PLoS One*, **7**, e42666.

75 Gaj, T., Mercer, A.C., Sirk, S.J., Smith, H.L. *et al.* (2013) A comprehensive approach to zinc-finger recombinase customization enables genomic targeting in human cells. *Nucleic Acids Res.*, **41**, 3937–3946.

76 Gaj, T., Mercer, A.C., Gersbach, C.A., Gordley, R.M. *et al.* (2011) Structure-guided reprogramming of serine recombinase DNA sequence specificity. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 498–503.

77 Gersbach, C.A., Gaj, T., Gordley, R.M., Mercer, A.C. *et al.* (2011) Targeted plasmid integration into the human genome by an engineered zinc-finger recombinase. *Nucleic Acids Res.*, **39**, 7868–7878.

78 Yant, S.R., Huang, Y., Akache, B., and Kay, M.A. (2007) Site-directed transposon integration in human cells. *Nucleic Acids Res.*, **35**, e50.

79 Kettlun, C., Galvan, D.L., George, A.L. Jr., Kaja, A. *et al.* (2011) Manipulating piggyBac transposon chromosomal integration site selection in human cells. *Mol. Ther.*, **19**, 1636–1644.

80 Wilson, M.H. and George, A.L. Jr. (2010) Designing and testing chimeric zinc finger transposases. *Methods Mol. Biol.*, **649**, 353–363.

81 Wilson, M.H., Kaminski, J.M., and George, A.L. Jr. (2005) Functional zinc finger/sleeping beauty transposase chimeras exhibit attenuated overproduction inhibition. *FEBS Lett.*, **579**, 6205–6209.

82 Li, X., Burnight, E.R., Cooney, A.L., Malani, N. *et al.* (2013) piggyBac transposase tools for genome engineering. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, E2279–E2287.

4

Rational Efforts to Streamline the *Escherichia coli* Genome

Gabriella Balikó, Viktor Vernyik, Ildikó Karcagi, Zsuzsanna Györfy, Gábor Draskovits, Tamás Fehér, and György Pósfai

Biological Research Centre of the Hungarian Academy of Sciences, Institute of Biochemistry, Synthetic and Systems Biology Unit, Temesvari krt. 62, Szeged, 6726, Hungary

Engineered biological parts, devices, and systems come to life when grafted into a living cell. Host cells are organisms shaped by billions of years of evolution and characterized by high complexity, robustness, and the ability to adapt and evolve in response to fluctuations in their natural environment. For synthetic biological applications, where precise engineering of biological systems with predictable outputs is attempted, host cells displaying reduced complexity, higher genetic stability, and increased efficiency are desired. We show here that streamlining, the elimination of genomic regions unnecessary or counterproductive in biotechnological applications, is a promising way to produce host cells, which can outperform their natural ancestors in the less fluctuating environment of laboratory settings. We focus here on the streamlining of *E. coli*, a primary host cell in research and industry. The rationale behind the streamlining process, identification of genomic parts targeted for elimination, deletion techniques, and results and applications of genome reduction projects will be presented. Current challenges, obstacles, and possible future directions of genome streamlining will also be discussed.

4.1 Introduction

Synthetic biological constructs – genetic circuits, modules, and devices – usually work in the context of a living cell. The information coded in the artificial blueprint, and embedded in the host genome, must be maintained and expressed by the cellular machinery of information processing. Ideally, the new construct functions in a predictable way and uses the cellular resources without much interference with the basic physiology of the host.

Natural host cells, even relatively simple bacteria, however, provide an extremely complex and frequently unpredictable environment for the synthetic construct, causing interference with the desired function [1, 2]. Moreover, since

Synthetic Biology: Parts, Devices and Applications, First Edition. Edited by Christina Smolke.

living cells possess the intrinsic ability for physiological and genetic adaptation, unwanted genotypic and phenotypic alterations may arise when challenged by artificial genetic constructs [3–5].

Conveniently, recent advances in genome manipulation and synthetic DNA construction techniques [6–11] as well as our rapidly expanding knowledge of the wealth of genome sequences [12] make genome-scale engineering possible, and, consequently, elimination of the disadvantageous features of the host cell can be attempted. Rationally redesigned, streamlined, and semisynthetic custom-made genomes could then replace naturally evolved gene sets, leading to an effective domestication of the microbial world [3, 11, 13–16].

In this chapter we will discuss the concept of the streamlined bacterial chassis, argue that *E. coli* is a primary choice for a versatile host, and review the tools and approaches of genome reduction. Next, results of *E. coli* genome streamlining and selected applications of the reduced-genome strains will be presented. Finally, future directions, gaps in our knowledge to be filled in, and perspectives of genome streamlining will be briefly discussed.

4.2 The Concept of a Streamlined Chassis

Natural cells are complex biological systems reflecting a long evolutionary history. The intrinsic functional robustness of natural cells, due to intertwining networks, functional redundancies, and feedback regulatory mechanisms make them resilient to synthetic reprogramming [17]. Moreover, their genomes are riddled with remnants of past adaptation events that may be irrelevant at present [18]. In addition, well-defined laboratory or industrial settings can be rather different from complex and changing natural environments [19, 20], rendering the existing genomic capabilities partially dispensable. Even if a number of empirically selected or purposefully introduced modifications shaped the genomes of some widely used experimental or industrial organisms, they still are unnecessarily complex and heterogeneous biological systems with a vast number of components and network interactions, largely unsuitable for precise and rational engineering.

Developing simple cells that provide only the very basic cellular machinery for maintaining and expressing designed constructs in a predefined range of conditions would thus greatly facilitate predictable engineering. Such a biological "chassis" could be used as a starting point to add new modules and build more complex systems adjusted to special needs [21]. Moreover, creating a more amenable and embraceable system would facilitate our understanding of general biological phenomena, such as transcriptome complexity, energy metabolism, and robustness [22, 23]. (It should be noted that a somewhat different interpretation of the chassis restricts it to a DNA-less cellular container, into which synthetic genomes could be transplanted [21]. We use here the term for a self-sustaining cellular system, complete with a simple genome.)

What are the desired features of a cellular chassis? First, it should have significantly reduced complexity. By eliminating unnecessary components, predictability of reprogramming could be enhanced. Second, the chassis should be

genetically stable. Even if mutagenesis and evolvability cannot be totally repressed, genetic change should be kept at the minimum to preserve the designed functionality. Third, by eliminating dispensable, energy-consuming components, the chassis should function more economically and utilize the resources efficiently, allowing high-yield product formation under well-defined conditions. In addition, the biological chassis should be safe for health and for the environment. By embedding genetic barriers in the blueprint, accidental release and genetic mixing with the natural organisms can be prevented.

Construction of a simple cell can be attempted in two ways. On one hand, building genomes from scratch, using synthetic oligonucleotide assemblies is an approach of great potential [3]. Despite the theoretical challenges of bottom-up genome design, the toolbox of genome assembly and transplantation into a living cell is undergoing continuous development [24–26]. The grandiose project of synthesizing a minimal genome (a genome comprising only essential genes), seen for *Mycoplasma mycoides*, may therefore become general practice one day [15]. On the other hand, rational simplification and optimization of existing robust cells in routine laboratory use is a less challenging and less risky endeavor. Beyond the elimination of unnecessary genes (streamlining), creating a chassis might involve other modifications as well: altering the genetic code (codon swaps), introduction of non-interfering subsystems (orthogonality), and redesign and rewiring (optimization) [6, 21]. Here we will discuss genome streamlining by focusing on the reduction of the *E. coli* genome.

Using the term genome streamlining we do not mean creating an absolute minimal set of genes required for life. Rather, the aim here is to produce a significantly reduced genome that retains all the important genes required for robust growth and easy genetic manipulation in a practical, laboratory, or industrial setting.

4.3 The *E. coli* Genome

E. coli is an important commensal and pathogen, an excellent model for research, and one of the most widely used industrial organisms. Among thousands of isolates, five strains (K-12, B, C, Crooks, and W) and their derivatives have been used extensively in laboratories for over 70 years [27]. Biotechnological applications range from production of commodity chemicals and biofuels to vaccine development and bioremediation. Notably, nearly 30% of approved recombinant therapeutic proteins are currently produced in *E. coli*.

Popularity of *E. coli* is owed to its versatility, simple culturability, and ease of genetic manipulation. *E. coli* can utilize a wide range of carbon and energy sources, is capable of aerobic growth and anaerobic fermentation, and can survive not only in the intestinal tract but also in the outside environment. The versatility of the bacterium is reflected in its relatively large (4.5–5.5 Mb) [28], high gene-density genome.

The genome sequence of the prototype laboratory strain K-12 MG1655 became available in 1997 [18] (selected features shown in Figure 4.1). The 4.6 Mb chromosome contains ~4300 protein-coding genes, accounting for about 88% of the genome. The remaining part encodes stable RNAs (0.8%) and

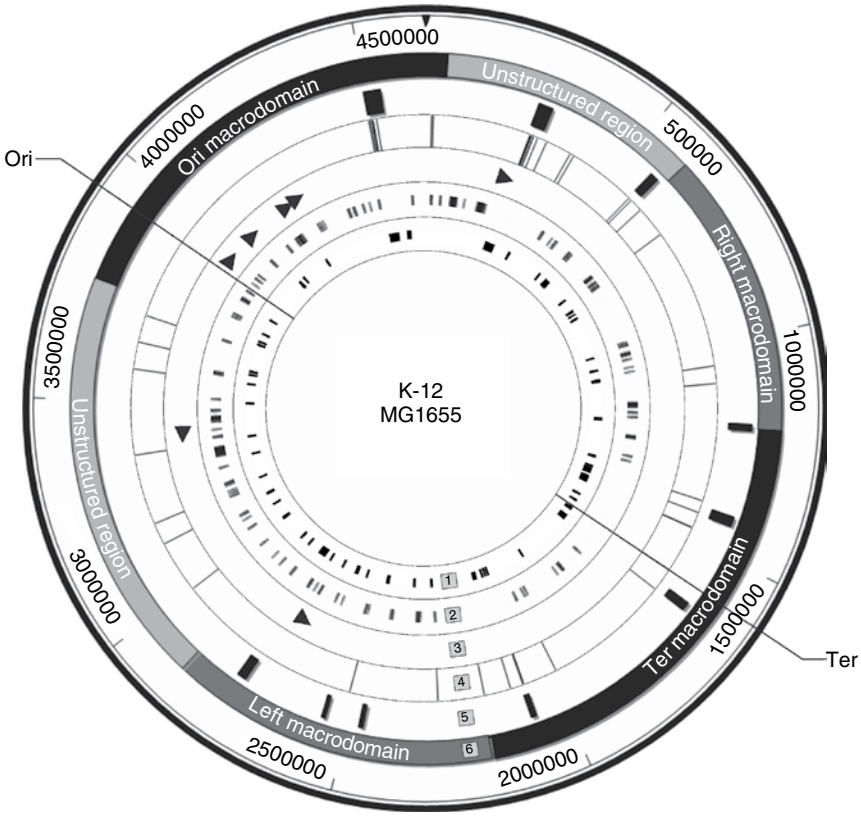


Figure 4.1 Schematic map of selected features of the *E. coli* K-12 MG1655 genome, numbered on the perimeter in base pair. Outward from the center, rings depict (1) strain-specific K-12 genomic islands longer than 4 kbp [29], (2) essential genes (www.shigen.nig.ac.jp/ecoli/pec/index.jsp), (3) ribosomal RNA operons, (4) IS elements, (5) prophages [18], and (6) macrodomains [30]. Ori and ter indicate the origin and terminus of replication, respectively.

provides regulatory and other functions (~11%). The genome is thus nearly fully loaded with information-bearing sequences, leaving very little room for apparently useless, intergenic DNA with no obvious function. The largest group of genes codes for transport and binding proteins, reflecting the wide variety of substrates the bacterium can utilize. Surprisingly, despite the long laboratory history of *E. coli*, 38% of the genes had no experimentally verified function at the time of sequencing, and even today this number stands about 20% [31].

As more genome sequences of *E. coli* strains became available, a peculiar, mosaic-like genome structure was revealed. The genomes share a common, homologous colinear backbone sequence, interrupted by hundreds of strain-specific genomic islands. Typically, these genomic islands carry marks of relatively recent horizontal transfer events and are characterized by a higher than average number of unknown genes, mobile genetic elements, and a relatively high A+T content. Since the basic cellular functions seem to be coded on the backbone sequences and, less importantly, life style-specific genes reside on

the genomic islands, they are called "core genome" and "auxiliary genome," respectively.

How many genes belong to the core genome? Obviously, the more genomes are compared, the smaller the core genome appears, and the core identified within a phylogroup is larger than the core obtained by inclusion of distant relatives. A comparison of 61 sequenced *E. coli* genomes revealed that out of a huge pan-genome of 15 741 gene families, only 993 (6%) of the families were represented in every genome (core genome) [32]. The accessory genes thus make up more than 90% of the pan-genome and about 80% of a typical genome [32]. It should be noted, however, that selection criteria applied to find conserved genes might miss homologs in distantly related strains. Moreover, alternative genetic solutions might exist for the same function. A refined comparison of 186 sequenced *E. coli* genomes [33], identifying homolog gene clusters (HGCs), revealed a pan-genome of 16 373 HGCs. The "soft core," defined as all HGCs found in at least 95% of the genomes, consisted of 3051 HGCs (Figure 4.2). A recent census, listing 2085 sequenced *E. coli* genomes, revealed that the pan-genome still grew linearly with the number of genomes added, while the size of the core genome of 3188 gene families hardly changed [34].

Why do we think that a significant part of the genome is dispensable without loss of fitness? *E. coli* evolved its gene set in the lower gut of animals, with periodic shedding in the environment. It has obviously many genes that are

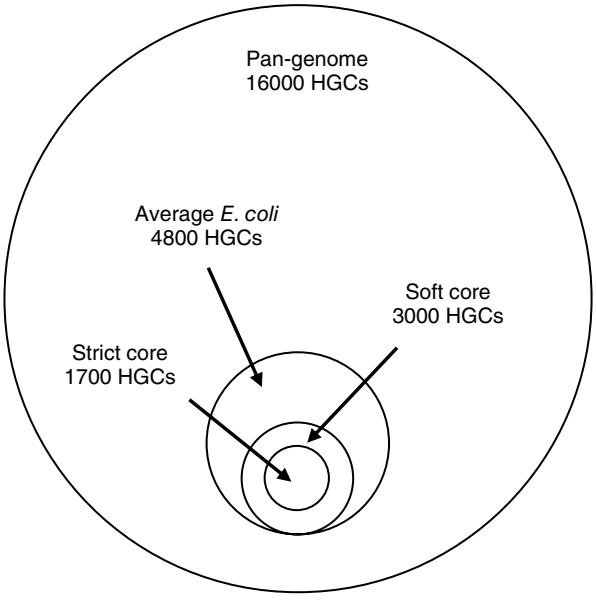


Figure 4.2 Comparison of the pan-genome and core-genome sizes, defined by homologous gene clusters (HGCs). Data and classification criteria are from [33] and are based on the analysis of 186 sequenced *E. coli* genomes. HGCs are generated by sequence similarity (95% of HGCs have <0.242 substitutions per site). The soft-core genome is defined as all HGCs that have members in at least 95% of the 186 genomes. The strict core genome is defined as all HGCs that have members in all genomes. The pan-genome is defined as all HGCs.

irrelevant under defined laboratory or industrial conditions. It was estimated that even under poor nutritional conditions, only 75–80% of the genes have detectable activity [35, 36]. Moreover, the genome is loaded with prophages and transposable elements (mostly residing on the accessory genome) (Figure 4.1), which, although occasionally contribute to fitness under certain conditions, could be viewed as dispensable genomic parasites. Finally, the fact that a large proportion of the genes lack a known function, despite of decades of *E. coli* research, suggests that they may be unimportant.

4.4 Random versus Targeted Streamlining

There are natural organisms possessing a nearly minimal number of genes, often in the range of 400–600. These organisms are typically obligate host-associated bacteria, and phylogenetic studies indicate that the small gene sets evolved from much larger genomes through massive loss of genes no longer required in the intracellular environment. This suggests that nutrient-rich, constant environment and low population size favor genome reduction. It was estimated that the free-living ancestor of *Buchnera* has lost 75% of its genome since it switched to an endosymbiotic lifestyle approximately 200 million years ago [37]. As an analogy, culturing a population of cells by serial passage under conditions favoring loss of genetic material (limiting nutrients for DNA synthesis, periodic population bottlenecks, defects in mismatch repair) could lead to smaller genomes [38–40]. Such an undirected procedure would have several advantages. First, no *a priori* knowledge of the genome is required. Second, high-fitness, rapidly growing cells are automatically selected. Third, this approach allows the exploration of different orders and combinations of deletion events. Unfortunately, since DNA synthesis requires little energy [22], there is no strong selection for smaller genome *per se*. Experimental work along this line so far has not resulted in major genome reduction. The 0.05–2.5 bp per genome per division deletion rate, obtained in an experimental evolution test with *Salmonella enterica* [41], is too low for practical application. Similarly, a long-term laboratory evolution experiment applying serial passage of *E. coli* cells in a single medium yielded only a few deletions totaling 38 kb in 20 000 generations [42]. Clearly, an experimental approach based on selectable deletion formation is needed for satisfactory results on a realistic time scale. An interesting approach partially fulfilled this requirement. Using an engineered, composite transposon, serial random deletions were created in *E. coli* [43]. Transposon-inserted cells were selected in each cycle by their antibiotic resistance. Subsequent induction of an "inner" transposon resulted in deletion (or inversion) of a neighboring genomic segment along with the loss of the resistance cassette, and a new cycle could be initiated. Unfortunately, there are some drawbacks: only one-fourth of the transposon-inserted cells undergo the proper rearrangement, replica plating is needed to find the proper clones, small deletions are favored, and the construct leaves a 64-bp exogenous sequence in the genome in each cycle. In conclusion, due to lack of an adequate deletion selection scheme, random deletion methods are currently not applied to genome streamlining. Instead, targeted genome reduction schemes are favored.

Rational, serial construction of targeted genomic deletions requires the full knowledge of the genome sequence, high quality gene annotations, sufficiently deep knowledge of cellular physiology, and adequate engineering tools. All these prerequisites are fulfilled for commonly used *E. coli* strains. Targeted, rational design has several advantages: there is no deletion size constraint *per se*, no subsequent identification of the modifications is required, and optimal serial strategy can be devised (subdivisions of deletions can be made and subsequently merged). Significantly, the process can be controlled at every step: in case a deletion causes an undesired effect (e.g., loss of fitness), the actual step can be skipped. On the other hand, the targeted approach suffers from historical contingency: cells with only predesigned deletions, introduced in an order of limited variability, are being created and tested.

4.5 Selecting Deletion Targets

4.5.1 General Considerations

It is not a trivial task to rationally select dispensable portions of the genome. The goal is to obtain a streamlined genome that still supports robust and rapid growth on a range of customary substrates. Usefulness of a gene, obviously, is context dependent, and our knowledge of the cellular and molecular network responses under dynamically changing environmental conditions is very limited. However, there are some gene categories that most likely represent negligible contribution to fitness under most conditions. There are several approaches that help identifying these targets.

4.5.1.1 Naturally Evolved Minimal Genomes

The small genomes of obligate symbionts and parasites can provide a template for a basic set of genes needed for maintaining cellular life. However, simply taking them as a blueprint for a simple organism can be misleading. Since essential nutrients and protection are usually provided by the host, the 400–600 genes they typically harbor are not sufficient to maintain life [13].

4.5.1.2 Gene Essentiality Studies

In most free-living organisms investigated, essential genes make up 10–30% of the genome. For *E. coli*, there are several large-scale gene essentiality studies available. High-throughput random transposon mutagenesis [44] or systematic gene inactivations [45] were applied to determine the subset of genes, which are indispensable. However, essentiality studies are not fail-proof. First, essentiality is a function of the environmental context. Second, both query methods might miss some hits. Transposon mutagenesis studies assume that a gene, which does not suffer an insertion event is essential, thus some genes escaping insertion by chance will be misqualified as essential. Moreover, single or grouped gene inactivations might not reveal redundant, but essential functions, and, conversely, might identify seemingly essential genes that, in fact, can be deleted in combination with other genes. Nevertheless, the 295 genes listed as essential candidates

("genes that have not been shown to be nonessential"; http://ecoliwiki.net/colipedia/index.php/Essential_genes 28 May 2013) (Figure 4.1) should obviously be retained in the streamlining process.

4.5.1.3 Comparative Genomics

Genome comparisons of related strains are highly informative and probably give the best clues as to what to delete. Natural selection, within the genus, supposedly conserved the basic set of genes, collectively called as the core genome, which are needed for robust performance [33]. Interspersed, horizontally acquired genomic islands carrying niche-specific and parasitic genes are obvious choices for removal (Figure 4.1). Although non-orthologous gene displacement might obscure shared functions [46], genome comparisons of more distantly related species could also help finding deletion targets. For example, *Buchnera* sp. is thought to be a naturally minimized version of *E. coli*, sharing a common ancestor before switching to a symbiotic lifestyle. The 0.64 Mbp genome of *Buchnera* could serve to identify genes common with *E. coli* and probably being important for growth. Genes unique to *E. coli* could then be used as a smaller pool to identify deletion candidates by other methods [47].

4.5.1.4 *In silico* Models

Genome-scale metabolic network reconstructions coupled with constraint-based modeling can contribute to rational strain design by predicting gene essentiality and phenotypic consequences of gene deletions in microbes. Although these large-scale computational models continue to be expanded and updated, their predictive power to quantitatively assess cellular phenotypes in streamlining studies is still limited [48]. In particular, these models often fail to identify groups of metabolic genes that are individually dispensable, but jointly essential. The most widely used *E. coli* reconstruction, while covering 1366 metabolic genes, still contains only a subset of the full gene complement of the cell [49]. Integration of other cellular systems (e.g., the machineries for replication, transcription, translation, posttranslational modifications) and regulatory processes is needed to more accurately compute complex cellular phenotypes [50, 51]. In addition, there are still too many unknown gene functions to accurately build an *in silico* interaction network that covers all key cellular processes.

4.5.1.5 Architectural Studies

Genome streamlining does not equal simply minimizing the gene set. The minimal set of genetic information necessary to sustain a functioning cell might contain positional information as well: not only trans-acting genes but also cis-acting chromosomal regions might be essential. In a comprehensive study [52], the entire chromosome was scanned for cis-acting regions. Essential genes were deleted from the chromosome in the presence of complementing plasmids carrying the particular gene. Surprisingly, the replication origin was found to be the only essential cis-acting region. Other, reportedly cis-acting regions, like *dif* (participating in resolution of replicated sister chromosomes) or *migS* (responsible for the polar movement of *oriC*) proved to be nonessential, and removal of them caused only minor growth defects.

In conclusion, genome streamlining is in large part a trial-and-error process. The large number of genes with unknown functions and the complex interactions of the constituents of the cell make precise *a priori* assessments difficult, especially when synergistic effects of serial deletions are considered. Nevertheless, based on the general considerations and on individual assessments, some gene categories can be marked as primary targets for deletion.

4.5.2 Primary Deletion Targets

4.5.2.1 Prophages

Strains of *E. coli* harbor multiple prophages or phage-related elements that may represent a significant fraction of the genome (typically 3–5%) (Figure 4.1). Prophages have a long history of coevolution with their host and seem to be well integrated in the host physiology. Typically, their genes code for integrases, lysozymes, and phage structural proteins, but they may carry metabolic and toxin–antitoxin functions as well. Compared with the entire genome, a higher than average number of prophage genes have no known function [18, 53]. Regarding their effect on the desired cell characteristics, prophages are Janus-faced. They can stimulate cell growth in certain conditions and can help the host to cope with a number of adverse conditions; however, under other conditions, their effect can be reduced growth, increased sensitivity [54], and instability [55]. Although most of the prophages are cryptic, normally unable to excise and develop infectious particles, some may excise and lyse the host upon stress [56]. Lytic phage development can be fatal for subsequent cultures of non-lysogenic strains that may be infected and destroyed [57]. Overall, removal of prophages and phage remnants does not seem to have adverse effects under customary growth conditions and may promote uniformity and stability of the culture.

4.5.2.2 Insertion Sequences (ISs)

Insertion sequences (ISs) are small mobile genetic elements carrying the minimal genetic information (inverted repeat ends and transposase gene) for their own genomic insertion [58]. Typically dozens of ISs of several different classes reside in the genomes of *E. coli* strains (Figure 4.1). ISs are important agents of genetic diversity and are responsible for a significant portion of the mutational load for the cell. While there are well-documented cases when ISs contribute to adaptation of the cell to specific conditions, they can generally be viewed as genomic parasites causing genetic instability, especially under stress [59]. From the practical perspective, removal of ISs significantly increases genetic stability without adverse effects. In fact, there are cases where presence of ISs prevents stabile cloning of toxic genes by mutagenesis and selection of altered clones [5].

4.5.2.3 Defense Systems

Common restriction systems of *E. coli* (*hsdMRS*, *mcrBC*, and *mrr*) and clustered regularly interspaced short palindromic repeat (CRISPR) systems provide defense against invasive foreign genetic material [60, 61]. While these systems are important factors in the interplay of evolutionary forces shaping the genomes,

they can be a nuisance in synthetic biology constructions. Deleting them eliminates barriers to genome engineering procedures that involve the introduction of genetic material into a heterologous host.

4.5.2.4 Genes of Unknown and Exotic Functions

A significant portion of the genome codes for genes with unknown function (~20%) [31]. Not excluding the possibility of discovering new and important functions, deletion of these genes might be attempted with high confidence. Similarly, metabolic and transport genes associated with substrates not commonly used are primary targets for removal. It might be intuitively argued that metabolic genes not needed under a particular condition are not expressed; hence deletion of them provides little gain in the economical use of resources. However, in fact, it was shown that, under conditions of declining carbon source quality, cells switch into a scavenging mode and express a variety of transport and metabolic genes to prepare for any substrate availability [62]. Thus, even if actually not used, exotic transport and metabolic genes can pose a metabolic burden on the cell.

4.5.2.5 Repeat Sequences

The largest repeat sequences of *E. coli*, rearrangement hot spot (Rhs) elements, are about 8 kb in length on average and collectively constitute about 1% of the genome [18]. Although widespread in *E. coli* strains, their function is poorly understood [63]. Rhs elements carry dispensable genes responsible for polysaccharide synthesis and export and for genes with unknown functions and might promote RecA-dependent rearrangements of the chromosome and are thus undesired for synthetic biology applications.

4.5.2.6 Virulence Factors and Surface Structures

The commonly used *E. coli* K-12 MG1655 strain is non-pathogenic due to the lack of a type-III secretion system and haemolysin expression, in addition to an impaired O-antigen synthesis [18]. Nevertheless, the strain harbors a number of virulence-associated factors, like flagella, fimbriae, siderophores and a cryptic haemolysin. There is a theoretical chance that safe strains acquire mutations or horizontally transferred additional virulence factors that transform them into a pathogen. It is a cause for concern that a double point mutation change in the gene coding for histon-like protein HUα can turn K-12 into an invasive strain [64]. Deletion of the genes associated with virulence therefore makes the cells safer. Elimination of surface structures might bring about other gains as well. For instance, the flagellar apparatus, not needed in a fermentor, consumes an estimated 1–2% of the total cellular energy. In addition, flagella break off and regrow constantly, and these proteins, shed in the environment, constitute a net loss for the cell [65]. Deletion of flagellar and chemotaxis gene clusters might thus result in energy savings. Elimination of other surface structures (fimbriae, curli, lipopolysaccharide outer core, colanic acid capsule) could further improve the cellular economy and also reduce the propensity of the cell for biofilm formation [66].

4.5.2.7 Genetic Diversity-Generating Factors

SOS-induced translesion DNA polymerases (polII, polIV, and polV) are major sources of mutations in the cell [67, 68]. When DNA is damaged, these polymerases rescue cells by bypassing bulky replication blocks and, at the same time, introduce point mutations in the genome. Whether the repair function or the generation of genetic diversity is the primary function is still debated. It seems that in case of moderate stress, alternative repair pathways can cope with the damage, but translesion polymerases are nevertheless induced and generate mutations [69, 70]. Deletion of the genes of translesion DNA polymerases is thus desirable to keep evolvability of the cell at the minimum. Indeed, it was shown that elimination of the translesion polymerases reduces the mutation rate of unstressed cells and, more significantly, prevents the increase of the mutation rate under stress. Engineered constructs, which pose a burden on cell growth and are therefore prone to deterioration via mutation and selection, can be maintained at higher fidelity in such a stabilized host [4]. It should be noted, however, that in case of heavy stress and DNA damage, when more extensive DNA repair is needed, lack of the translesion polymerases may cause a reduction in fitness [70].

4.5.2.8 Redundant and Overlapping Functions

There are several redundant or overlapping functions in *E. coli*, and deletion of some of them can be attempted presumably without compromising growth and robustness. Typical examples include DNAses, RNAses, and transport systems. For instance, quadruple and quintuple mutations of nucleases were applied, albeit at a fitness cost under certain conditions, in order to increase the stability of electroporated oligonucleotides, enhancing the efficiency of oligonucleotide-mediated allelic replacement procedures [71, 72].

4.6 Targeted Deletion Techniques

4.6.1 General Considerations

E. coli is usually viewed as one of the most readily amenable organism for genetic engineering, with an arsenal of genetic engineering tools available. However, not all *E. coli* strains can be equally well manipulated by the usual tools. Differences in restriction and recombination systems, variable transformation efficiency and antibiotic sensitivity, resistance to transducing phage, and restricted applicability of the counterselecting sacB–sucrose system are a few examples of potential obstacles. From the engineering point of view, K-12 derivatives are the best-suited strains. To date, nearly all serial, large-scale *E. coli* genome streamlining projects have been performed in such cell lines.

Construction of targeted, base pair precision deletions is usually based on homologous recombination of dsDNA. To create a deletion, an engineered DNA segment, carrying a selection marker and sequences matching the flanking genomic regions of a desired deletion, is transformed in the cell, where exchange with the genomic segment takes place, catalyzed by endogenous recombinases.

In a second recombination event, the exogenous sequences can be excised to leave a markerless deletion. Repeating the steps, multiple deletions can be cumulated in the cell [13].

A semi-random genome reduction attempt, combining deletions derived from mapped transposon-inserted genomic libraries, applied site-specific recombinase systems (Flp/frt, Cre/lox) for the excision step [73]. The procedure, however, leaves a scar, a 34-bp recognition site in the genome, which may interfere with subsequent rounds of deletions. The problem can be circumvented by the use of mutant recognition sites, but the scheme is complex, and still a scar is left behind. Most large-scale streamlining projects therefore used improved, general homology-based deletion methods, producing scarless deletions.

Mutant target sites of site-specific recombinases, preimplanted in the genome, can also be used to facilitate the exchange of long DNA fragments between an episome and the chromosome. Using this strategy, a 126 kbp-long chromosomal segment was replaced with a 72 kbp synthetic DNA cassette carrying three noncontiguous genomic deletions. The subsequent elimination of the remaining loxP sites by homologous recombination and introduction of novel mutant loxP sites can in theory make this somewhat complicated process applicable for large-scale genome reduction [74].

4.6.2 Basic Methods and Strategies

4.6.2.1 Circular DNA-Based Method

Suicide plasmids, replicons multiplying only under permissive conditions, serve as delivery vehicles for deletion-forming DNA constructs [75, 76]. The plasmid carries fused homology arms (~0.5–1.0 kb long DNA segments matching the flanking sequences of the genomic region to be deleted) for the first recombination event, an antibiotic resistance gene as a selection marker, and a gene (usually *sacB*) allowing counterselection [77] in the second recombination step. Integration into the genome at one side of the planned deletion occurs via recombination between one of the homology arms and the corresponding chromosomal sequence, catalyzed by RecA. Such co-integrates are selected by their antibiotic resistance under nonpermissive conditions for plasmid replication. Next, cells that resolved the co-integrate in a spontaneous, second recombination event, are selected applying counterselection procedures (e.g., permitting growth of only $sacB^-$ cells by using sucrose-containing medium [78, 79]). Outcome of the procedure can be either recovery of wild type or formation of a scarless deletion. An advanced, more effective version of the method applies I-SceI cleavage (the enzyme cuts the co-integrate at a 18-bp site [80] found exclusively in the integrated sequence) as a universal counterselection tool, which, at the same time, stimulates recombination [81] (Figure 4.3a).

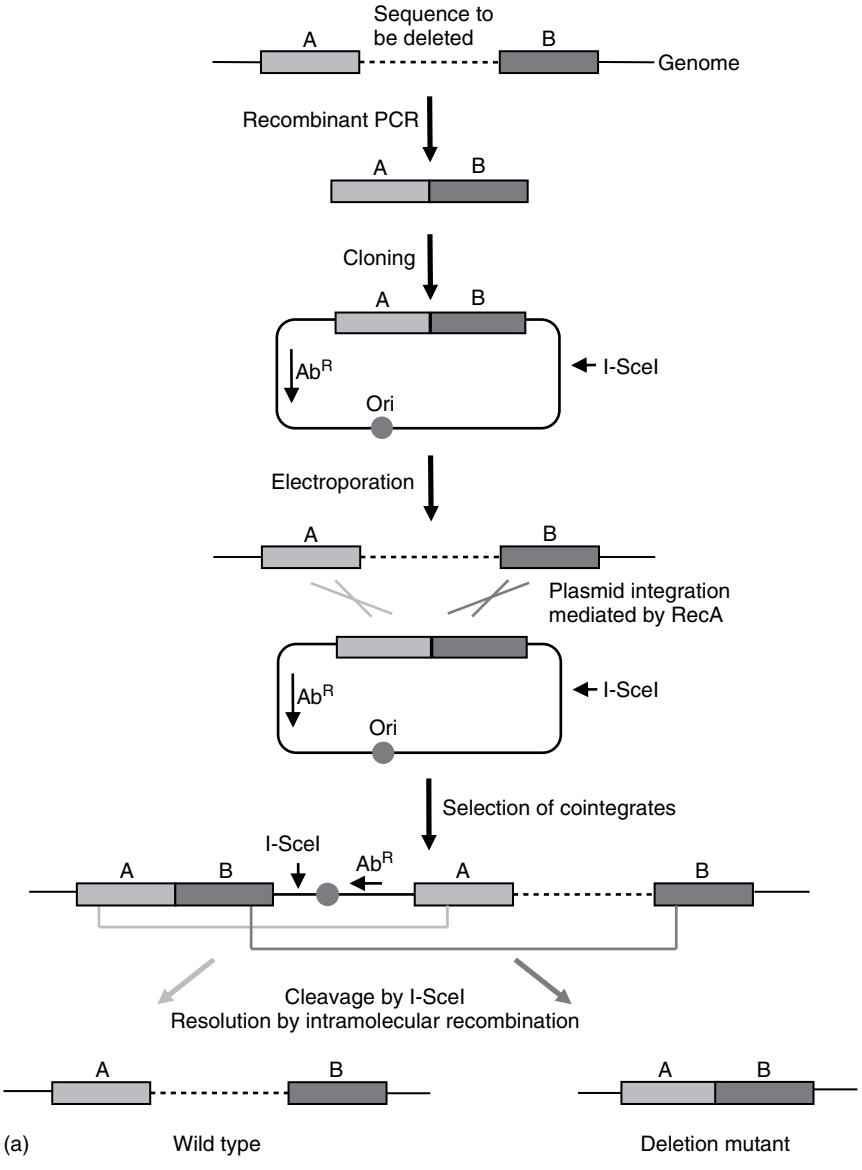


Figure 4.3 General scheme of standard deletion procedures. (a) Overview of the circular DNA-based method. Boxes A and B represent >500-bp DNA segments flanking the genomic region to be deleted. AbR stands for an antibiotic resistance marker gene; ori indicates a replication origin functioning only under permissive conditions. (b) Overview of the λ-Red-mediated, linear DNA-based deletion method. Two alternative routes for generating deletions are shown. A, B, and C represent arbitrarily chosen 40–60-bp DNA segments (homology boxes). Arrowheads represent I-SceI cleavage sites. Ab^R and csm stand for an antibiotic resistance marker and a counterselectable gene, respectively.

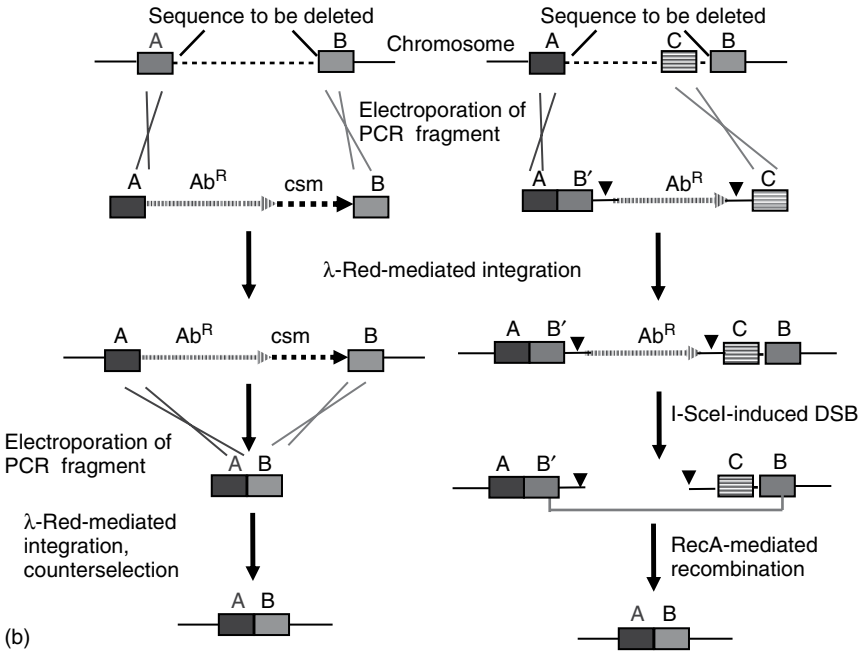


Figure 4.3 (*Continued*)

4.6.2.2 Linear DNA-Based Method

A more straightforward method applies a polymerase chain reaction (PCR)-generated linear dsDNA fragment carrying flanking homology arms and genes for selection and counterselection (Figure 4.3b). Recombination into the genome is catalyzed by the lambda Red system [82], expressed typically from a plasmid. The second recombination step, using a linear DNA fragment composed of the flanking homology arms, is applied to replace the exogenous sequences, creating a scarless deletion. Homology arms can be as short as 40 bp [83, 84] but work best when longer (1 kb; [47]). The method is straightforward, and even large deletions (~100 kb) can be obtained efficiently. Incorporating a third homology box and a I-SceI cleavage site in the targeting dsDNA fragment alleviates the need for recombination of a second targeting fragment, accelerating the procedure. Induced I-SceI cleavage of the integrated sequence stimulates intramolecular recombination between the third homology box and a matching neighboring genomic sequence, resulting in a scarless deletion [85](Figure 4.3b).

4.6.2.3 Strategy for Piling Deletions

Accumulating deletions in a cell one by one is a labor-intensive endeavor. Some simple strategical considerations help accelerating the process. Individual deletion intermediates (e.g., unresolved co-integrates carrying a selection marker) can be made and checked for fitness in a parallel fashion. Genomic segments carrying the selected deletion intermediates can then be sequentially transferred into the multiple deletion strain by cycles of P1 transduction and

I-SceI-stimulated scarless resolution. In principle, addition of new deletions to the final host can be accomplished in a multiplex, iterative fashion. This might allow the combination of the best deletion candidates, selected due to faster growth.

4.6.2.4 New Variations on Deletion Construction

Several overlapping studies demonstrate an approach which couple the CRISPR/Cas9 system with λ-Red-mediated recombineering [86–90]. In contrast to previous strategies, it does not rely on chromosomal integration and subsequent removal of selectable markers. Since *E. coli* lacks the nonhomologous end-joining (NHEJ) repair system, double-stranded chromosomal breaks are highly lethal, unless rescued by providing a bridging template DNA segment. This strategy requires targeted double-stranded DNA cleavage by Cas9 and λ-Red-mediated genomic integration of a homologous template DNA carrying the desired deletion. The donor DNA can be either single or double stranded and might be introduced as a plasmid or in a linear form. Chromosomal cleavage not only facilitates recombination but also provides strong counterselection against the wild-type cells; therefore the efficiency of this tool can be very high, up to 100%.

CRISPR/Cas9-derived nickases were also used to generate targeted deletions between genomic repeats [91]. They showed that creating single-stranded chromosomal incisions by mutant Cas9 nucleases are not lethal; moreover, it facilitates the intramolecular recombination between repetitive elements. Dual-targeted nicking in IS element repeats generated two deletions in one step, removing a total of 133 kbp from the genome.

The CRISPR/Cas9 coupled with NHEJ system from mycobacteria enables rapid and continuous creation of large deletions without applying selection markers or homologous DNA template [92, 93]. First, CRISPR/Cas9-targeted double-stranded breaks are generated flanking the desired deletion. Next, the NHEJ proteins seal the DNA ends in an imprecise way and thus rescue the cells. Using this powerful technique, deletion of a 123 kbp genomic fragment was demonstrated [93].

Another way of using CRISPR/Cas nucleases to facilitate λ-Red-mediated genome editing is to provide long linear DNA fragments by cleaving bacterial artificial chromosomes (BACs) *in vivo*. Both the BAC cleavage and the genomic recombination processes are selected for using appropriately placed positive/negative selection markers. This method, referred to as replicon excision for enhanced genome engineering through programmed recombination (REXER), has been used to replace a 230 kbp-long genomic segment of *E. coli* and could be a promising technique for the stepwise re-coding of the complete chromosome [16].

A related strategy referred to as multiple essential genes assembling (MEGA) applies the I-SceI endonuclease to release a linear DNA fragment from a circular plasmid *in vivo* [94]. The released fragment comprises all essential genes corresponding to the targeted genomic region. Subsequent cleavage of the I-SceI sites inserted into the chromosome generates a double-strand break that facilitates the replacement of the genomic region with the essential gene cluster by λ-Red recombination.

Recently, rapid streamlining and genome-wide inactivation of IS elements were accomplished by genome shuffling between different *E. coli* strains, followed by multiplex genome modifications by CRISPR/Cas-assisted MAGE [95]. First, prophages were deleted by shuffling prophage-free segments of multiple deletion series (MDS) genomes into *E. coli* BL21 by P1 transduction. This was followed by subsequent rounds of CRISPR/Cas-assisted MAGE on multiplex IS targets, disrupting the transposases of the IS elements. With the growing number of reduced-genome strains, such recycling of streamlined genomes might accelerate strain construction.

4.7 Genome-Reducing Efforts and the Impact of Streamlining

4.7.1 Comparative Genomics-Based Genome Stabilization and Improvement

The first systematic, large-scale genome reduction project was aimed at removing the largest K12-specific genomic islands from the MG1655 genome [85] (Figure 4.4). Identification of the K-islands was based on the sequence comparison of three *E. coli* genomes available at that time (MG1655, enterohemorrhagic O157:H7, and uropathogenic CFT073). Via a series of linear DNA-mediated recombineering steps, including a novel way of I-SceI-stimulated scarless resolution of the recombination intermediate, 12 precise deletions were created and combined in a single strain. Compared with the parental MG1655, the resulting MDS12 (multiple deletion series strain with 12 deletions) had a genome reduced by 8.1%, with 9.3% of the genes deleted. All prophages and 24 of the 44 transposable elements present in the MG1655 genome were deleted. Growth rates of MDS12 in minimal and rich medium were similar to those of MG1655. Doubling times were nearly identical, but MDS12 reached 10% higher density in stationary phase. Electroporation and transformation efficiencies of the parental and the MDS12 strain were identical. This first attempt of drastic genome streamlining proved that by applying a rational design strategy, a large fraction of the genes can be removed from an organism that has been shaped by billions of years of evolution. Moreover, this could be done without losing robustness and rapid growth, at least under the laboratory conditions tested.

The next milestones of this project were the IS-free MDS41, MDS42, and MDS43 strains with 14.28, 14.30, and 15.27% of the genome deleted, respectively [29]. Deletion targets were primarily selected by comparative genomics of several sequenced strains (RS218, CFT073, *Shigella flexneri* 2457T, O157:H7 EDL933, and DH10B) and by assessment of literature data on the particular gene functions. Major K-islands were targeted, but deletions were in several cases extended to include neighboring nonessential genes with no impact on growth in either rich or minimal media. Deletions were tested for growth properties both individually and when combined in a single strain. Growth rates of the MDS cells were similar to that of the parental MG1655. Elimination of

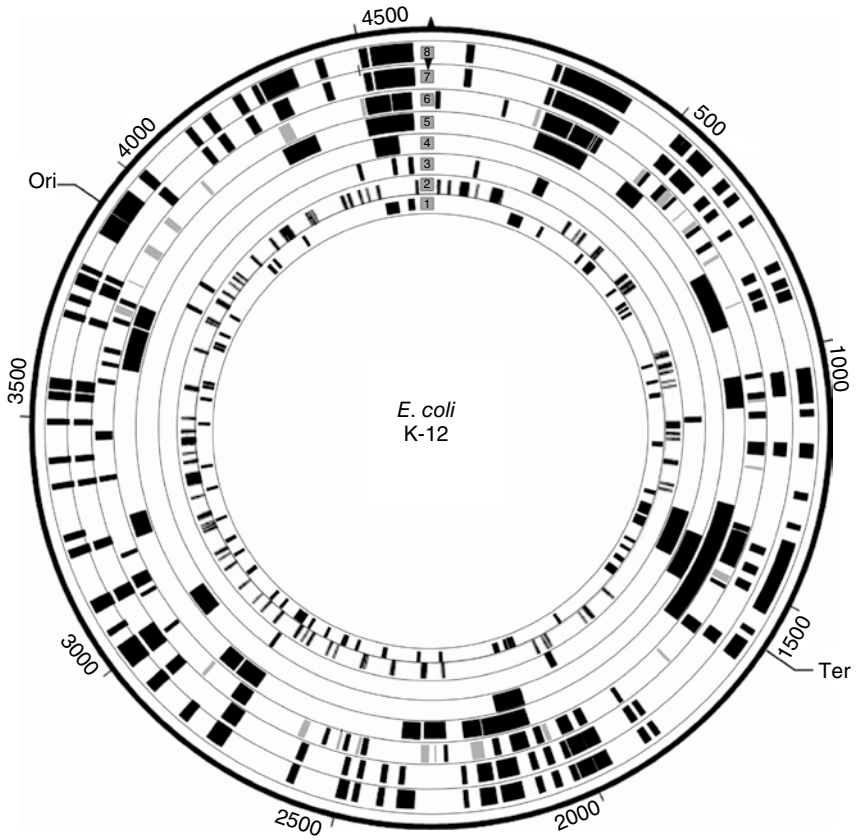


Figure 4.4 Deletion map of reduced-genome *E. coli* strains. Rings depict features mapped to the genome of *E. coli* K-12 MG1655, numbered on the perimeter in kilobase pair. Outward from the center, (1) strain-specific K-12 genomic islands longer than 4 kbp [96], (2) essential genes (www.shigen.nig.ac.jp/ecoli/pec/index.jsp), and (3)–(8) set of deletions constructed by Goryshin *et al*. [43], Yu *et al*. [73], Hashimoto *et al*. [97], Pósfai *et al*. (MDS42: black boxes, MDS69: black and gray boxes) [29, 85], Mizoguchi *et al*. (MGF-01) [47], and Hirokawa *et al*. (DGF-298) [98], respectively. Ori and ter indicate the origin and terminus of replication, respectively.

recombinogenic or mobile DNA stabilized the MDS genomes and provided a host free of IS contamination for plasmid preparations and gene libraries, reducing the chances for cloning artifacts, solving a frequently arising but usually overlooked problem. Many cryptic virulence genes were also removed, presumably increasing the safety of the strains. High yields of recombinant protein production were achieved in MDS cells. Genome reduction also led to unanticipated beneficial properties: high electroporation efficiency and accurate propagation of recombinant genes and plasmids with strong secondary structure that were unstable in other strains. It was demonstrated that the stability of lentiviral vectors containing long direct repeats was significantly enhanced in MDS42 [99]. Genome stability was further increased by deleting the three SOS-inducible, error-prone DNA polymerases PolII, PolIV, and PolV [4], significantly reducing point-mutation rates, thereby allowing more faithful expression

of a heterologous toxic protein. Versions of MDS42, rendering it less recombinogenic (recA$^-$), suitable for blue-white selection cloning (lacZΔM15), or expressing inducible T7 polymerase are also available for common applications. Continuing the MDS series, reduced-genome strains with up to 69 deletions were created by removing further putatively horizontally transferred regions [100].The final member of the series, MDS69 lost 965 genes, 20.3% of the genome.

A novel, rapid streamlining workflow, the MDS series based on genome shuffling and CRISPR/Cas-assisted MAGE was developed to improve the stability of the *E. coli* BL21(DE3), a host frequently used for high-level recombinant protein production [101]. All 9 resident prophages were deleted and all 50 active IS elements were removed or inactivated. The DE3 prophage carrying an inducible T7 RNA polymerase gene was exchanged with a tightly controlled T7 RNA polymerase cassette. Additional strain variants with inactivated error-prone DNA polymerases were also constructed. The streamlined BL21(DE3)-K-12 hybrid strains retained the favorable characteristics of BL21(DE3), displayed increased genomic and plasmid stability, and allowed elevated electroporation efficiencies [95].

4.7.2 Genome Reduction Based on Gene Essentiality

In another attempt to reduce the genome of *E. coli* MG1655, a series of medium-scale and large-scale markerless deletions were constructed using linear targeting, DNA/λ-Red-mediated recombination, and *sacB*-based counterselection methods (Figure 4.4) [97]. Deletion targets were selected by excluding essential genes and maximizing the potentially deletable chromosomal segments. First, many nonessential regions were removed from the chromosome. Next, by combining consecutive deletions, a series of mutants were constructed, lacking up to 29.7% (16 combined deletions) of the chromosome. Mutants with individual deletions grew like the wild-type strain. The mutants with an increasing number of combined deletions, however, grew increasingly slower than the parental strain in rich medium. The mutant with the largest number of deletions (16) grew much slower than the parental strain (45.4 min vs. 26.2 min doubling time) and showed aberrant nucleoid morphology, as well as altered cell shape and size. It was concluded that the additive effect of large deletions can sometimes not be predicted, but the deletion of nonessential chromosome regions may be valuable for elucidating cellular processes governed by multiple systems.

The interspersed nature of essential genes within bacterial chromosomes would normally require genome reduction projects to execute numerous short deletions targeting the flanked nonessential segments. To accelerate this process, the MEGA (see above) technique replaces long chromosomal stretches with short DNA cassettes comprising solely the essential genes of the targeted segment [94]. As a proof of principle, three regions ranging from 80 to 205 kbp were deleted this way in the *E. coli* chromosome with each target containing two to eight essential genes. The authors envisioned the stepwise, complete replacement of the *E. coli* genome with the gene set essential to sustain life.

4.7.3 Complex Streamlining Efforts Based on Growth Properties

In a study focusing on cell growth in minimal medium, long, scarless deletions of *E. coli* W3110 were constructed (Figure 4.4) [47]. The long-term goal of the work is to create streamlined-genome strains, which are suitable platforms for metabolic engineering. To identify deletable chromosomal segments, the genome sequences of *E. coli* K-12 MG1655 and *Buchnera* sp. APS were used for comparative genomics, and genes unique for *E. coli* were selected. Essential genes reported in the PEC database (http://www.shigen.nig.ac.jp/ecoli/pec/index.jsp) were excluded from the deletion list. The annotations of the remaining genes were surveyed in databases to judge their importance for efficient growth in M9 minimal medium, and regions with more than 10 continuous deletable genes were chosen for deletion. Genomic deletions were made by using λ-Red-mediated recombination and the negative selection marker *sacB*. Individual deletion strains were checked for growth in M9 minimal medium, and only the well-growing constructs were chosen for further use. By combining the individual deletions, a top-performer strain (designated minimum genome factory 01 (MGF-01)) with 1 Mb total genome reduction was obtained. MGF-01 grew as fast as the parental W3110 strain and reached higher optical density and higher number of colony-forming units (CFUs) in stationary phase in minimal medium. This higher-density growth property emerged by superpositioning the individual deletions and might be caused by the lower level accumulation of growth-inhibiting acetate, presumably due to the elevated expression of glyoxylate shunt-related genes *aceA* and *aceB*. This more efficient metabolism could also be the reason MGF-01 with an L-threonine-producing unit integrated into the genome produced 2.4-fold higher amount of L-threonine than the parental strain carrying the same unit.

The genome of MGF-01 was further reduced via step-by-step accumulation of additional deletions made in W3110 (Figure 4.4) [98]. Noncore regions were chosen for deletions. Starting with 37 individual deletions, strains with normal phenotypes were selected, and 10 of them were added to MGF-01 in subsequent cycles, generating MGF-02. Analysis of the growth phenotype of MGF-02 revealed that deletion of *gcvA* encoding a positive regulator of the glycine cleavage system enhanced initial growth in minimal medium. To further optimize the strain, two intrinsic mutations of parental MG1655, *ilvG* and *rph-1* (causing valine sensitivity and partial pyrimidine starvation, respectively), were fixed both in MGF-01 and MGF-02, creating DGF-362 and DGF-348, respectively. Starting from DGF-348, further deletions were added by keeping only those without growth-reducing synergistic effects. The *proVXW* carrying region, deleted initially, was reintroduced into the genome to fix sensitivity to high osmolarity. Eventually, the strain with the smallest genome (DGF-298) possessed a 2.98 Mb chromosome and was free from all IS elements. DGF-298 grew better in M9 minimal medium than parental W3110 and also had higher cell yield in a simple medium (CSL) in fermentation. Transcriptome analysis showed that a heat-shock chaperone (IbpAB) and a protease for abnormal proteins (Lon) are downregulated in DGF strains. The authors concluded that downregulation of the genes encoding chaperones and proteases is one of the factors that improve the fitness of DGF strains.

4.7.4 Additional Genome Reduction Studies

In an early proof-of-concept work, random genomic deletions were constructed in MG1655 by developing a method involving repeated integration/deletion of a Tn5 transposon derivative. Deleted regions could be rescued on a conditionally replicating plasmid, allowing identification of essential genes. The extent of the genome reduction in the most deleted strain was about 5.6%, as estimated by pulsed-field electrophoresis [43] (Figure 4.4).

Utilizing a pre-mapped transposon insertion library, a semi-random method was applied to reduce the genome of MG1655 by 6.7% (Figure 4.4) [73]. A pair of selected transposon insertions could be combined in a single cell by P1 transduction, and the genomic region between them could be excised by the Cre/lox system. Combining of deletions in a single genome was also achieved by P1 transduction. In some multiple deleted strains, synthetic lethality was observed: some deletions were individually viable but were lethal when combined. This genome engineering strategy, producing large sets of mapped transposon insertions ready for pairwise combination, followed by Cre/lox-mediated in between deletion, is most useful when deletion of a particular region of the genome is desired.

4.8 Selected Research Applications of Streamlined-Genome *E. coli*

4.8.1 Testing Genome Streamlining Hypotheses

The MDS series with increasing number of genomic deletions provides a convenient model for studying the impact of stepwise genome streamlining on cellular traits, addressing unsettled questions of reductive genome evolution [100]. A comprehensive study showed that deletions caused a gradual fitness loss, decreased nutrient utilization, and induced a general stress response. Growth yield and maintenance energy were measured in chemostat cultures of MG1655, MDS42, and MDS69 under nutrient limitation. Both carbon and nitrogen utilization efficiencies decreased in the multideletion strains without significantly affecting the maintenance energy requirement of the cell. These results argue against the adaptive genome streamlining hypothesis [102, 103]. Results supported the notion that selection for reduced DNA synthesis per se is unlikely to reduce genome size in the course of evolution of small genomes. No general trend was found between growth rate and genome size, neither between cell size and genome size. Genome reduction was also shown to cause transcriptome reprogramming. Many targets of the general stress sigma factor RpoS were upregulated in MDS42 and MDS69. rprA, a small regulatory RNA that facilitates RpoS translation was strongly induced, and, as expected, the MDS42 and MDS69 had elevated acid resistance. These studies revealed an unexpectedly significant role of horizontally transferred genes not only in stressful environments but also under routine growth conditions.

4.8.2 Mobile Genetic Elements, Mutations, and Evolution

Bacterial genomes are usually loaded with a great number of ISs of many types. The evolutionary forces driving their accumulation and their general impact on adaptive evolution of the host are unknown. IS-free MDS42 provides a unique opportunity to investigate the initial spread and evolutionary impact of ISs. By introducing a single IS1 element into the genome MDS42, its impact could be analyzed in laboratory evolutionary experiments. Although the IS element increased the mutational supply and contributed to adaptation, another mutator gene (*mutS*), frequently found in natural isolates, had a much greater impact on the evolution of the cell. Moreover, *mutS* cells outcompeted IS-carrying cells, limiting their spread. This work showed that the initial spread of IS elements might depend on the presence of other mutator mechanisms in the population, hence demonstrating the evolutionary conflict between different mutation-generating mechanisms [104].

Mobile element-free strains were also used in synthetic biology studies. To improve the stability of synthetic genetic circuits, bidirectional (overlapping forward and backward) promoters were designed to couple transcription of a target nonessential gene to the transcription of an essential gene. The evolutionary half-life of the gene of interest increased 4–10 times, and the circuit was more stable in the IS-free MDS42 than in MG1655. However, eventually point mutations, insertions/deletions and recombination occurred even in MDS42, demonstrating the need for further stabilization of synthetic constructs [105].

4.8.3 Gene Function and Network Regulation

MDS42 proved to be especially useful in transcriptional studies elucidating the physiological role and the molecular mechanisms of the rho-dependent transcription termination system [54]. Rho silences foreign DNA, repressing prophages and other horizontally acquired portions of the genome, but this function becomes less important in MDS42 that lacks prophages and many horizontally transferred regions. As a consequence, MDS42 shows 10^4 times lower sensitivity to the Rho-inhibitor bicyclomycin than the ancestor MG1655. Moreover, Rho cofactors NusA and NusG, normally essential in *E. coli*, become dispensable in MDS42.

Reduced-genome strains were used to identify the genes required for biofilm development [106]. They found new genes, some of them being cryptic in MG1655 but expressed in the reduced-genome mutant, discovered by this approach. In addition, by means of the deletion strains, a new repressor was identified for starvation-sensing protein RspA [107].

The relationship between the genomic and environmental contributions to the transcriptome was analyzed by comparing the transcriptomes of MG1655 and MDS42 grown in regular and transient heat-shock conditions. Results suggest a cross-talk guiding transcriptional reorganization in *E. coli* in response to both genetic and environmental disturbances [108].

4.8.4 Codon Reassignment

Incorporation of unnatural amino acids (uaas) in proteins in living cells would enable evolution of novel protein functions [109]. Relatively rarely used stop codons, coupled with orthogonal tRNA/synthetase pairs, can be exploited to genetically introduce uaas. A major limitation of using a stop codon to encode uaas is the low efficiency of incorporation due to competition of the suppressor tRNA with endogenous release factors (RF1 and RF2 in prokaryotes). UAG is the least used stop codon in *E. coli* (present in 7% of the genes) and is recognized by RF1, but not by RF2. To achieve full reassignment of UAG, the reportedly essential RF1 must be removed from the system. It was shown that, after modifying the activity of RF2, the gene encoding RF1 (*prfA*) can be deleted from the *E. coli* genome. MDS42 was used as parental strain, because the deletion of nearly 700 genes may alleviate the termination load imposed on RF2. Besides the demonstrated successes for multisite incorporations of uaas for protein research and laboratory evolution, the RF1 knockout strains can also be valuable for investigating the evolution of the genetic code [110].

Due to the degenerate nature of the genetic code, reassigning sense codons to encode uaas is also conceivable, once the specific codons are successfully eliminated from the genome. In a proof-of-concept work, the synonymous re-coding of certain Ser, Leu, or Ala codons was attempted in a 20 kbp-long essential operon of *E. coli* MDS42 [16]. Eight different re-coding schemes were tested, some of which resulted in the exchange of 373 codons in a single step. Measuring the efficiency of various codon exchanges permitted the definition of allowed and disallowed synonymous re-coding schemes to be applied in future codon reassignment projects.

A similar project, on the long run, aimed at the re-coding of the complete *E. coli* MDS42 genome to eliminate all 62 214 instances of seven different codons [111]. In this endeavor, re-coding would take place by the stepwise exchange of 50 kbp-long segments of the chromosome. Testing the complementing ability of the synthetic recoded DNA segments one by one, 99.5% of the recoded genes were found to complement their wild-type counterparts without the need of further optimization. The use of the MDS42 strain in such re-coding enterprises warrants reduced synthesis costs and improved genome stability.

4.8.5 Genome Architecture

As reduced-genome bacteria have altered positions and perturbed local context of certain chromosomal segments, these strains could be useful for studying genome architectural effects. A comparative protein occupancy profile of MG1655 and MDS42 was analyzed using microarray-based chromatin immunoprecipitation [112]. This work identified both highly transcribed and transcriptionally silent extended protein occupancy domains, *hi*EPODs and *ts*EPODs, respectively. It was suggested that the binding of *ts*EPODs by nucleoid proteins (HU, Fis, H-NS, and IHF) establishes them as chromosomal organizing centers. MDS42 lacks a large fraction of *ts*EPODs, but the remaining ones are similarly located as in parental MG1655, supporting a dynamic role of the organizing centers in the formation of a higher-order chromosome structure.

4.9 Concluding Remarks, Challenges, and Future Directions

Streamlined-genome *E. coli* strains are representatives of a promising direction of synthetic biology research. The goals of cell simplification have already been partially fulfilled. Reduced complexity arising from elimination of redundant and unnecessary functions helped to elucidate hitherto unknown functions and network interactions. Increased phenotypic uniformity and genetic stability can be exploited for maintaining unstable synthetic constructs. Demonstration of increased amino acid production by reduced-genome cells may hint at improvements in cellular economy.

Numerous applications, from bacterial computation and gene network model building to vaccine production and plasmid biopharmaceutical manufacturing, have been suggested for streamlined-genome *E. coli*. However, most tangible applications to date were research oriented, and despite all the advances, published biotechnological applications of streamlined-genome cells were limited to a few pilot studies (Table 4.1). In order to attain a more widespread status as production hosts, simplified cells clearly need improvements, and superior performance over traditional production strains have to be demonstrated.

On one hand, construction of a superior chassis should involve not only streamlining but also extensive rational optimization. Introduction of mutations known to increase fitness or compensating for loss of certain genetic material could enhance performance. Advantageous features of different *E. coli* strains (e.g., the high recombinant protein production capability of BL21 and the easy genetic accessibility and high stability of K-12 MDS) could be combined in a single host [95]. New genome manipulation techniques are at hand to accelerate the optimization process. MAGE allows simultaneous, targeted introduction of small modifications at many genomic sites [121]. New DNA

Table 4.1 Published biotechnology-related applications of streamlined-genome *E. coli*.

Application	References
Recombinant protein production	[113–115]
Construction of lentiviral expression vectors	[99]
Enhanced L-threonine production	[47, 116]
Stabilized maintenance of genetic constructs	[4, 5]
Expression of avian influenza virus gene	[117]
Dengue reporter virus constructions	[118]
Periplasmic delivery of human interleukin-10	[119]
Investigation of antimicrobial peptide sensitivity	[120]
Construction of IS-free P1 phage	[7]
Incorporation of unnatural amino acids in proteins	[110]

cleavage tools (TALENS, CRISPS/Cas), tailored to target specific genomic sites, might be used to devise novel schemes to rapidly perform various manipulations [8, 122–124].

On the other hand, rational, targeted streamlining, and optimization could be complemented by random engineering coupled with directed evolution. Devising efficient, forced random deletion-creating schemes, applying cyclic multiplex genomic alteration techniques, or shuffling different genomes would vastly increase the number of genomic variants, from which the fittest versions could be identified by proper selection.

The plummeting cost of DNA synthesis is continuously increasing the relevance and reality of synthesizing streamlined genomes. Originally, bottom-up synthesis and top-down reduction of genomes were viewed as two competing and opposite approaches to simplify bacterial cells. In current practice, these two strategies seem to harmonically complement each other: reduced genomes are used as starting points of complete genetic rewiring using synthetic DNA cassettes [16, 111], and deletion construction has also been demonstrated with plasmids carrying synthetic DNA fragments [74]. Furthermore, the boundary of the two strategies is blurred *ab ovo*, for the gene sets of minimal genomes synthesized to date are all subsets of the genetic repertoire of extant bacterial strains [15]. It is possible, however, that in the future, minimal genomes will be synthesized by combining genes originating from multiple species.

How far should genome size reduction extend? In general, the relatively small effects of the extensive genomic perturbations represented by streamlined genomes attest to a remarkable robustness of the cellular physiology and genome architecture. However, reduced complexity inevitably comes at the expense of robustness and adaptability to external factors [23]. Observations from practical genome streamlining works ([97] and our observations) also suggest that large-scale elimination of genes, while initially resulting in improvements, may reduce robustness and cause deterioration of basic cellular physiology (growth properties, adaptability, nucleoid structure, cell morphology) beyond a certain point that roughly corresponds to the core-genome size (Figure 4.5).

Beyond the complexity issue, physical constraints on genome size might also limit reduction efforts. Despite decades of research, little is known on the homeostatic mechanisms coordinating DNA replication, transcription, and translation to maintain a constant DNA to cell mass ratio [125]. Significantly altering the genome size may perturb these mechanisms. Moreover, while our knowledge regarding gene and network functions is getting even more complete, constraints of genome architecture *per se* are less understood [126]. The specific and relative localization of some genes (e.g., ribosomal RNA operons) and specific chromosomal sites (e.g., binding sites for proteins participating in cell division), superhelicity of the genome, and macro- and microdomain structure are all influenced in a largely unknown way by genome reduction.

However, synthetic biology provides us just the appropriate tools to address these issues [9]: streamlined genomes can be specifically designed and constructed to elucidate the constraints of genome size and architecture.

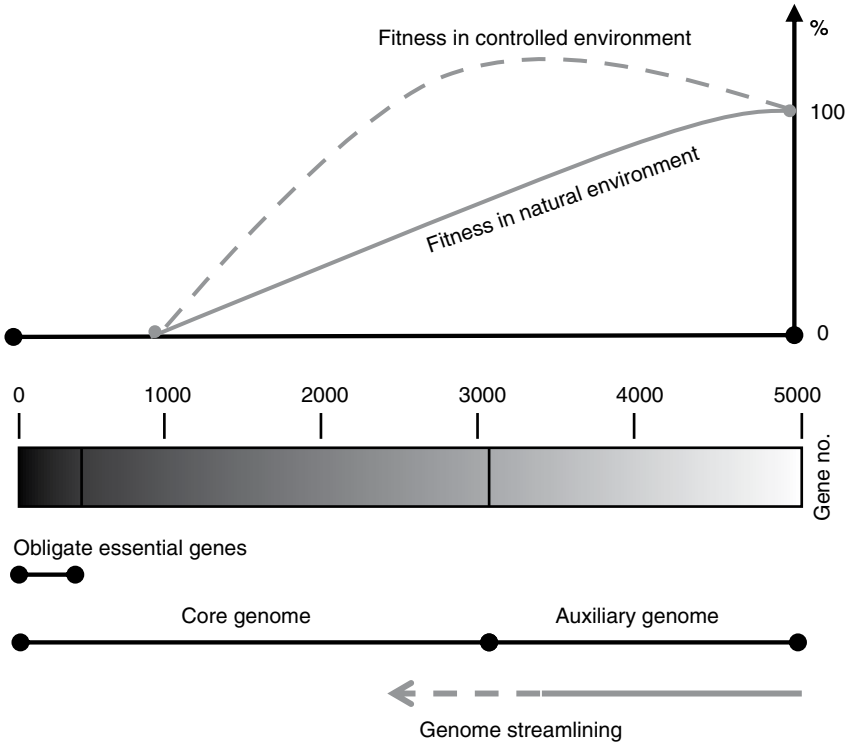


Figure 4.5 Hypothetical relationship between the fitness of the cell and the extent of genome streamlining.

References

1 Arkin, A. (2008) Setting the standard in synthetic biology. *Nat. Biotechnol.*, **26**, 771–774.

2 Moon, T.S., Lou, C., Tamsir, A., Stanton, B.C. *et al.* (2012) Genetic programs constructed from layered logic gates in single cells. *Nature*, **491**, 249–253.

3 Gibson, D.G., Glass, J.I., Lartigue, C., Noskov, V.N. *et al.* (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. *Science*, **329**, 52–56.

4 Csorgo, B., Feher, T., Timar, E., Blattner, F.R. *et al.* (2012) Low-mutation-rate, reduced-genome *Escherichia coli*: an improved host for faithful maintenance of engineered genetic constructs. *Microb. Cell Fact.*, **11**, 11.

5 Umenhoffer, K., Feher, T., Balikó, G., Ayadin, F. *et al.* (2010) Reduced evolvability of *Escherichia coli* MDS42, an IS-less cellular chassic for molecular and synthetic biology applications. *Microb. Cell Fact.*, **9**, 38.

6 Esvelt, K.M. and Wang, H.H. (2013) Genome-scale engineering for systems and synthetic biology. *Mol. Syst. Biol.*, **9**, 1–17.

7 Feher, T., Karcagi, I., Blattner, F.R., and Posfai, G. (2012) Bacteriophage recombineering in the lytic state using the lambda red recombinases. *Microb. Biotechnol.*, **5**, 466–476.

8 Csorgo, B., Nyerges, A., Posfai, G., and Feher, T. (2016) System-level genome editing in microbes. *Curr. Opin. Microbiol.*, **33**, 113–122.

9 Pal, C., Papp, B., and Posfai, G. (2014) The dawn of evolutionary genome engineering. *Nat. Rev. Genet.*, **15**, 504–512.

10 Selle, K. and Barrangou, R. (2015) Harnessing CRISPR-Cas systems for bacterial genome editing. *Trends Microbiol.*, **23**, 225–232.

11 Martinez-Garcia, E. and de Lorenzo, V. (2016) The quest for the minimal bacterial genome. *Curr. Opin. Biotechnol.*, **42**, 216–224.

12 Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G. *et al.* (2017) Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic Acids Res.*, **45**, D446–D456.

13 Fehér, T., Papp, B., Pál, C., and Gy, P. (2007) Systematic genome reductions: theoretical and experimental approaches. *Chem. Rev.*, **107**, 3498–3513.

14 Feller, S.M. (2010) Life v2.0 – Quo vadis Homo sapiens? *Cell Commun. Signaling*, **8**, 9.

15 Hutchison, C.A. 3rd, Chuang, R.Y., Noskov, V.N., Assad-Garcia, N. *et al.* (2016) Design and synthesis of a minimal bacterial genome. *Science*, **351**, aad6253.

16 Wang, K., Fredens, J., Brunner, S.F., Kim, S.H. *et al.* (2016) Defining synonymous codon compression schemes by genome recoding. *Nature*, **539**, 59–64.

17 Vickers, C.E., Blank, L.M., and Kromer, J.O. (2010) Grand challenge commentary: chassis cells for industrial biochemical production. *Nat. Chem. Biol.*, **6**, 875–877.

18 Blattner, F.R., Plunkett, G.I., Block, C.A., Perna, N.T. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1462.

19 Almirón, M., Link, A.J., Furlong, D., and Kolter, R. (1992) A novel DNA-binding protein with regulatory and protective roles in starved *E. coli*. *Genes Dev.*, **6**, 2646–2654.

20 Kolter, R., Siegele, D.A., and Tormo, A. (1993) The stationary phase of the bacterial cell cycle. *Annu. Rev. Microbiol.*, **47**, 855–874.

21 Acevedo-Rocha, C.G., Fang, G., Schmidt, M., Ussery, D.W. *et al.* (2013) From essential to persistent genes: a functional approach to constructing synthetic life. *Trends Genet.*, **29**, 273–279.

22 Wodke, J.A., Puchalka, J., Lluch-Senar, M., Marcos, J. *et al.* (2013) Dissecting the energy metabolism in *Mycoplasma pneumoniae* through genome-scale metabolic modeling. *Mol. Syst. Biol.*, **9**, 653.

23 Yus, E., Maier, T., Michalodimitrakis, K., van Noort, V. *et al.* (2009) Impact of genome reduction on bacterial metabolism and its regulation. *Science*, **326**, 1263–1268.

24 Karas, B.J., Jablanovic, J., Sun, L., Ma, L. *et al.* (2013) Direct transfer of whole genomes from bacteria to yeast. *Nat. Methods*, **10**, 410–412.

25 Noskov, V.N., Karas, B.J., Young, L., Chuang, R.Y. *et al.* (2012) Assembly of large, high G+C bacterial DNA fragments in yeast. *ACS Synth. Biol.*, **1**, 267–273.

26 Zhou, J., Wu, R., Xue, X., and Qin, Z. (2016) CasHRA (Cas9-facilitated homologous recombination assembly) method of constructing megabase-sized DNA. *Nucleic Acids Res.*, **44**, e124.

27 Archer, C.T., Kim, J.F., Jeong, H., Park, J.H. *et al.* (2011) The genome sequence of *E. coli* W (ATCC 9637): comparative genome analysis and an improved genome-scale reconstruction of *E. coli*. *BMC Genomics*, **12**, 9.

28 Bergthorsson, U. and Ochman, H. (1998) Distribution of chromosome length variation in natural isolates of *Escherichia coli*. *Mol. Biol. Evol.*, **15**, 6–16.

29 Pósfai, G., Plunkett, G., Fehér, T., Frisch, D. *et al.* (2006) Emergent properties of reduced-genome *Escherichia coli*. *Science*, **312**, 1044–1046.

30 Valens, M., Penaud, S., Rossignol, M., and Cornet, F. (2004) Macrodomain organization of the *Escherichia coli* chromosome. *EMBO J.*, **23**, 4330–4341.

31 Keseler, I.M., Collado-Vides, J., Santos-Zavaleta, A., Peralta-Gil, M. *et al.* (2011) EcoCyc: a comprehensive database of *Escherichia coli* biology. *Nucleic Acids Res.*, **39**, D583–D590.

32 Lukjancenko, O., Wassenaar, T.M., and Ussery, D.W. (2010) Comparison of 61 sequenced *Escherichia coli* genomes. *Microb. Ecol.*, **60**, 708–720.

33 Kaas, R.S., Friis, C., Ussery, D.W., and Aarestrup, F.M. (2012) Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. *BMC Genomics*, **13**, 577.

34 Land, M., Hauser, L., Jun, S.R., Nookaew, I. *et al.* (2015) Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics*, **15**, 141–161.

35 Tao, H., Bausch, C., Richmond, C., Blattner, F.R. *et al.* (1999) Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media. *J. Bacteriol.*, **181**, 6425–6440.

36 Wei, Y., Lee, J.M., Smulski, D.R., and LaRossa, R.A. (2001) Global impact of sdiA amplification revealed by comprehensive gene expression profiling of *Escherichia coli*. *J. Bacteriol.*, **183**, 2265–2272.

37 Moran, N.A. and Mira, A. (2001) The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. *Genome Biol.*, **2**, RESEARCH0054.

38 Andersson, S.G. and Kurland, C.G. (1998) Reductive evolution of resident genomes. *Trends Microbiol.*, **6**, 263–268.

39 Lovett, S.T. and Feschenko, V.V. (1996) Stabilization of diverged tandem repeats by mismatch repair: evidence for deletion formation via a misaligned replication intermediate. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 7120–7124.

40 Maniloff, J. (1996) The minimal cell genome: "on being the right size". *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 10004–10006.

41 Nilsson, A.I., Koskiniemi, S., Eriksson, S., Kugelberg, E. *et al.* (2005) Bacterial genome size reduction by experimental evolution. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 12112–12116.

42 Barrick, J.E., Yu, D.S., Yoon, S.H., Jeong, H. *et al.* (2009) Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature*, **461**, 1243–1247.

43 Goryshin, I.J., Naumann, T.A., Apodaca, J., and Reznikoff, W.S. (2003) Chromosomal deletion formation system based on Tn5 double transposition: use for making minimal genomes and essential gene analysis. *Genome Res.*, **13**, 644–653.

44 Gerdes, S.Y., Scholle, M.D., Campbell, J.W., Balazsi, G. *et al.* (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.*, **185**, 5673–5684.
45 Baba, T., Ara, T., Hasegawa, M., Takay, Y. *et al.* (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.*, **2**, 0008.
46 Koonin, E.V. (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.*, **1**, 127–136.
47 Mizoguchi, H., Sawano, Y., Kato, J., and Mori, H. (2008) Superpositioning of deletions promotes growth of *Escherichia coli* with a reduced genome. *DNA Res.*, **15**, 277–284.
48 McCloskey, D., Palsson, B.O., and Feist, A.H. (2013) Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*. *Mol. Syst. Biol.*, **9**, 661.
49 Feist, A.M., Herrgard, M.J., Thiele, I., Reed, J.L. *et al.* (2009) Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.*, **7**, 129–143.
50 Herrgard, M.J., Covert, M.W., and Palsson, B.O. (2004) Reconstruction of microbial transcriptional regulatory networks. *Curr. Opin. Biotechnol.*, **15**, 70–77.
51 Thiele, I., Jamshidi, N., Fleming, R.M., and Palsson, B.O. (2009) Genome-scale reconstruction of *Escherichia coli's* transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS Comput. Biol.*, **5**, e1000312.
52 Kato, J. and Hashimoto, M. (2007) Construction of consecutive deletions of the *Escherichia coli* chromosome. *Mol. Syst. Biol.*, **3**, 7.
53 Casjens, S. (2003) Prophages and bacterial genomes: what have we learned so far? *Mol. Microbiol.*, **49**, 277–300.
54 Cardinale, C.J., Washburn, R.S., Tadigotla, V.R., Brown, L.M. *et al.* (2008) Termination factor Rho and its cofactors NusA and NusG silences foreign DNA in *E-coli*. *Science*, **320**, 935–938.
55 Czyz, A., Los, M., Wrobel, B., and Wegrzyn, G. (2001) Inhibition of spontaneous induction of lambdoid prophages in *Escherichia coli* cultures: simple procedures with possible biotechnological applications. *BMC Biotech.*, **1**, 1.
56 Wang, X.X., Kim, Y., Ma, Q., Hong, S.H. *et al.* (2010) Cryptic prophages help bacteria cope with adverse environments. *Nat. Commun.*, **1**, 147.
57 Los, M., Czyz, A., Sell, E., Wegrzyn, A. *et al.* (2004) Bacteriophage contamination: is there a simple method to reduce its deleterious effects in laboratory cultures and biotechnological factories? *J. Appl. Genet.*, **45**, 111–120.
58 Siguier, P., Filee, J., and Chandler, M. (2006) Insertion sequences in prokaryotic genomes. *Curr. Opin. Microbiol.*, **9**, 526–531.
59 Lynch, M. (2007) *The Origins of Genome Architecture*, Sinauer Press, Sunderland, MA.
60 Bhaya, D., Davison, M., and Barrangou, R. (2011) CRISPS-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annu. Rev. Genet.*, **45**, 273–297.

61 Raleigh, E.A. (1992) Organization and function of the mcrBC genes of *Escherichia coli* K-12. *Mol. Microbiol.*, **6**, 1079–1086.

62 Liu, M., Durfee, T., Cabrera, J.E., Zhao, K. *et al.* (2005) Global transcriptional programs reveal a carbon source foraging strategy by *Escherichia coli*. *J. Biol. Chem.*, **280**, 15921–15927.

63 Jackson, A.P., Thomas, G.H., Parkhill, J., and Thomson, N.R. (2009) Evolutionary diversification of an ancient gene family (*rhs*) through C-terminal displacement. *BMC Genomics*, **10**, 584.

64 Koli, P., Sudan, S., Fitzgerald, D., Adhya, S. *et al.* (2011) Conversion of commensal *Escherichia coli* K-12 to an invasive form via expression of mutant histone-like protein. *mBio*, **2**, e00182-11.

65 Smith, D.R. and Chapman, M.R. (2010) Economical evolution: microbes reduce the synthetic cost of extracellular proteins. *mBio*, **1**, e00131-10–e00131-18.

66 Sung, B.H., Lee, C.H., Yu, B.J., Lee, J.H. *et al.* (2006) Development of a biofilm production-deficient *Escherichia coli* strain as a host for biotechnological applications. *Appl. Environ. Microbiol.*, **72**, 3336–3342.

67 Napolitano, R., Janel-Bintz, R., Wagner, J., and Fuchs, R.P. (2000) All three SOS-inducible DNA polymerases (Pol II, Pol IV and Pol V) are involved in induced mutagenesis. *EMBO J.*, **19**, 6259–6265.

68 Tippin, B., Pham, P., and Goodman, M.F. (2004) Error-prone replication for better or worse. *Trends Microbiol.*, **12**, 288–295.

69 Berdichevsky, A., Izhar, L., and Livneh, Z. (2002) Error-free recombinational repair predominates over mutagenic translesion replication in *E. coli*. *Mol. Cell*, **10**, 917–924.

70 Yeiser, B., Pepper, E.D., Goodman, M.F., and Finkel, S.E. (2002) SOS-induced DNA polymerases enhance long-term survival and evolutionary fitness. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 8737–8741.

71 Bethany, E.D., Sutera, V.A., and Lovett, S.T. (2007) RecA-independent recombination is efficient but limited by exonucleases. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 216–221.

72 Sawitzke, J.A., Thomason, L.C., Costantino, N., Bubunenko, M. *et al.* (2007) Recombineering: in vivo genetic engineering in *E. coli*, *S. enterica*, and beyond. *Methods Enzymol.*, **421**, 171–199.

73 Yu, B.J., Sung, B.H., Koob, M.D., Lee, C.H. *et al.* (2002) Minimization of the *Escherichia coli* genome using a Tn*5*-targeted Cre/*loxP* excision system. *Nat. Biotechnol.*, **20**, 1018–1023.

74 Krishnakumar, R., Grose, C., Haft, D.H., Zaveri, J. *et al.* (2014) Simultaneous non-contiguous deletions using large synthetic DNA and site-specific recombinases. *Nucleic Acids Res.*, **42**, e111.

75 Hamilton, C.M., Aldea, M., Washburn, B.K., Babitzke, P. *et al.* (1989) New method for generating deletions and gene replacement in *Escherichia coli*. *J. Bacteriol.*, **171**, 4617–4622.

76 Leenhouts, K., Buist, G., Bolhuis, A., ten Berge, A. *et al.* (1996) A general system for generating unlabelled gene replacements in bacterial chromosomes. *Mol. Gen. Genet.*, **253**, 217–224.

77 Reyrat, J.M., Pelicic, V., Gicquel, B., and Rappuoli, R. (1998) Counterselectable markers: untapped tools for bacterial genetics and pathogenesis. *Infection Immun.*, **66**, 4011–4017.

78 Gay, P., LeCoq, D., Steinmetz, M., Berkelman, T. *et al.* (1985) Positive selection procedure for entrapment of insertion sequence element in gram-negative bacteria. *J. Bacteriol.*, **164**, 918–921.

79 Link, A.J., Philips, D., and Church, G.M. (1997) Methods for generating precise deletions and insertions in the genome of wild-type *Escherichia coli*: application to open reading frame characterization. *J. Bacteriol.*, **179**, 6228–6237.

80 Monteilhet, C., Perrin, A., Thierry, A., Colleaux, L. *et al.* (1990) Purification and characterization of the *in vitro* activity of I-*Sce*I, a novel and highly specific endonuclease encoded by a group I intron. *Nucleic Acids Res.*, **18**, 1407–1413.

81 Pósfai, G., Kolisnychenko, V., Bereczki, Z., and Blattner, F.R. (1999) Markerless gene replacement in *Escherichia coli* stimulated by a double-strand break in the chromosome. *Nucleic Acids Res.*, **27**, 4409–4415.

82 Datsenko, K.A. and Wanner, B.L. (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 6640–6645.

83 Muyrers, J.P., Zhang, Y., Buchholz, F., and Stewart, A.F. (2000) RecE/RecT and Redalpha/Redbeta initiate double-stranded break repair by specifically interacting with their respective partners. *Genes Dev.*, **14**, 1971–1982.

84 Yu, D., Ellis, H.M., Lee, E.C., Jenkins, N.A. *et al.* (2000) An efficient recombination system for chromosome engineering in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 5978–5983.

85 Kolisnychenko, V., Plunkett, G.I., Herring, C.D., Fehér, T. *et al.* (2002) Engineering a reduced *Escherichia coli* genome. *Genome Res.*, **12**, 640–647.

86 Jiang, W., Bikard, D., Cox, D., Zhang, F. *et al.* (2013) RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.*, **31**, 233–239.

87 Jiang, Y., Chen, B., Duan, C., Sun, B. *et al.* (2015) Multigene editing in the *Escherichia coli* genome via the CRISPR-Cas9 system. *Appl. Environ. Microbiol.*, **81**, 2506–2514.

88 Li, Y., Lin, Z., Huang, C., Zhang, Y. *et al.* (2015) Metabolic engineering of *Escherichia coli* using CRISPR-Cas9 meditated genome editing. *Metab. Eng.*, **31**, 13–21.

89 Pyne, M.E., Moo-Young, M., Chung, D.A., and Chou, C.P. (2015) Coupling the CRISPR/Cas9 system with lambda red recombineering enables simplified chromosomal gene replacement in *Escherichia coli*. *Appl. Environ. Microbiol.*, **81**, 5103–5114.

90 Reisch, C.R. and Prather, K.L.J. (2017) Scarless Cas9 assisted recombineering (no-SCAR) in *Escherichia coli*, an easy-to-use system for genome editing. *Curr. Protoc. Mol. Biol.*, **117**, 31 8 1–31 8 20.

91 Standage-Beier, K., Zhang, Q., and Wang, X. (2015) Targeted large-scale deletion of bacterial genomes using CRISPR-nickases. *ACS Synth. Biol.*, **4**, 1217–1225.

92 Su, T., Liu, F., Gu, P., Jin, H. *et al.* (2016) A CRISPR-Cas9 assisted Non-homologous End-joining strategy for one-step engineering of bacterial genome. *Sci. Rep.*, **6**, 37895.

93 Zheng, X., Li, S.Y., Zhao, G.P., and Wang, J. (2017) An efficient system for deletion of large DNA fragments in *Escherichia coli* via introduction of both Cas9 and the non-homologous end joining system from Mycobacterium smegmatis. *Biochem. Biophys. Res. Commun.*, **485**, 768–774.

94 Xue, X., Wang, T., Jiang, P., Shao, Y. *et al.* (2015) MEGA (multiple essential genes assembling) deletion and replacement method for genome reduction in *Escherichia coli*. *ACS Synth. Biol.*, **4**, 700–706.

95 Umenhoffer, K., Draskovits, G., Nyerges, A., Karcagi, I. *et al.* (2017) Genome-wide abolishment of mobile genetic elements using genome shuffling and CRISPR/Cas-assisted MAGE allows the efficient stabilization of a bacterial chassis. *ACS Synth. Biol.*, **6**, 1471–1483.

96 Perna, N.T., Plunkett, G. 3rd, Burland, V., Mau, B. *et al.* (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*, **409**, 529–533.

97 Hashimoto, M., Ichimura, T., Mizoguchi, H., Keyamura, K. *et al.* (2005) Cell size and nucleoid organization of engineered *Escherichia coli* cells with a reduced genome. *Mol. Microbiol.*, **55**, 13.

98 Hirokawa, Y., Kawano, H., Tanaka-Masuda, K., Nakamura, N. *et al.* (2013) Genetic manipulations restored the growth fitness of reduced-genome *Escherichia coli*. *J. Biosci. Bioeng.*, 1–7.

99 Chakiath, C.S. and Esposito, D. (2007) Improved recombinatorial stability of lentiviral expression vectors using reduced-genome *Escherichia coli*. *Biotechniques*, **43**, 466–470.

100 Karcagi, I., Draskovits, G., Umenhoffer, K., Fekete, G. *et al.* (2016) Indispensability of horizontally transferred genes and its impact on bacterial genome streamlining. *Mol. Biol. Evol.*, **33**, 1257–1269.

101 Terpe, K. (2006) Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems. *Appl. Microbiol. Biotechnol.*, **72**, 211–222.

102 Cavalier-Smith, T. (2005) Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. *Ann. Bot.*, **95**, 147–175.

103 Giovannoni, S.J., Cameron Thrash, J., and Temperton, B. (2014) Implications of streamlining theory for microbial ecology. *ISME J.*, **8**, 1553–1565.

104 Fehér, T., Bogos, B., Méhi, O., Fekete, G. *et al.* (2012) Competition between transposable elements and mutator genes in bacteria. *Mol. Biol. Evol.*, **29**, 3153–3159.

105 Yang, S., Sleight, S.C., and Sauro, H.M. (2013) Rationally designed bidirectional promoter improves the evolutionary stability of synthetic genetic circuits. *Nucleic Acids Res.*, **41**, e33.

106 May, T. and Satoshi, O. (2011) Enterobactin is required for biofilm development in reduced-genome *Escherichia coli*. *Environ. Microbiol.*, **13**, 3149.

107 Sakihama, Y., Mizoguchi, H., Oshima, T., and Ogasawara, N. (2012) YdfH identified as a repressor of rspA by the use of reduced genome *Escherichia coli* MGF-01. *Biosci. Biotechnol., Biochem.*, **76**, 1688–1693.

108 Ying, B.W., Seno, S., Kaneko, F., Matsuda, H. *et al.* (2013) Multilevel comparative analysis of the contribution of genome reduction and heat shock to the *Escherichia coli* transcriptome. *BMC Genomics*, **14**, 25.

109 Liu, C.C. and Schultz, P.G. (2010) Adding new chemistries to the genetic code. *Annu. Rev. Biochem.*, **79**, 413–444.
110 Johnson, D.B., Xu, J., Shen, Z., Jeffrey, K. *et al.* (2011) RF1 knockout allows ribosomal incorporation of unnatural amino acids at multiple sites. *Nat. Chem. Biol.*, **7**, 779–786.
111 Ostrov, N., Landon, M., Guell, M., Kuznetsov, G. *et al.* (2016) Design, synthesis, and testing toward a 57-codon genome. *Science*, **353**, 819–822.
112 Vora, T., Hottes, A.K., and Tavazole, S. (2009) Protein occupancy landscape of a bacterial genome. *Mol. Cell*, **35**, 247–253.
113 Ran, H., Wu, J., Wu, D., and Duan, X. (2016) Enhanced production of recombinant *Thermobifida fusca* isoamylase in *Escherichia coli* MDS42. *Appl. Biochem. Biotechnol.*, **180**, 464–476.
114 Sharma, S.S., Blattner, F.R., and Harcum, S.W. (2007) Recombinant protein production in an *Escherichia coli* reduced genome strain. *Metab. Eng.*, **9**, 133–141.
115 Sharma, S.S., Campbell, J.W., Frisch, D., Blattner, R.F. *et al.* (2007) Expression of recombinant chloramphenicol acetyltransferase variants in highly reduced *Escherichia coli* strains. *Biotechnol. Bioeng.*, **98**, 1056–1070.
116 Lee, J.H., Sung, B.H., Kim, M.S., Blattner, F.R. *et al.* (2009) Metabolic engineering of a reduced-genome strain of *Escherichia coli* for l-threonine production. *Microb. Cell Fact.*, **8**, 2.
117 Gohrbandt, S., Veits, J., Breithaupt, A., Hundt, J. *et al.* (2011) H9 avian influenza reassortant with engineered polybasic cleavage site displays a highly pathogenic phenotype in chicken. *J. Gen. Virol.*, **92**, 1843–1853.
118 Schoggins, J.W., Dorner, M., Feulner, M., Imanaka, N. *et al.* (2012) Dengue reporter viruses reveal viral dynamics in interferon receptor-deficient mice and sensitivity to interferon effectors in vitro. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 14610–14615.
119 Poehlmann, C., Brandt, M., Mottok, D.S., Zschuettig, A. *et al.* (2012) Periplasmic delivery of biologically active human interleukin-10 in *Escherichia coli* via a sec-dependent signal peptide. *J. Mol. Microbiol. Biotechnol.*, **22**, 1–9.
120 Zschuettig, A., Zimmermann, K., Blom, J., Goesmann, A. *et al.* (2012) Identification and characterization of microcin S, a new antibacterial peptide produced by probiotic *Escherichia coli* G3/10. *PLoS One*, **7**, e33351.
121 Wang, H.H., Isaac, F.J., Carr, P.A., Sun, Z.Z. *et al.* (2009) Programming cells by multiple genome engineering and accelerated evolution. *Nature*, **460**, 894–898.
122 Cong, L., Ran, F.A., Cox, D., Lin, S. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.
123 Wood, A.J., Lo, T.W., Zeitler, B., Pickle, C.S. *et al.* (2011) Targeted genome editing across species using ZFNs and TALENs. *Science*, **333**, 307.
124 Tas, H., Nguyen, C.T., Patel, R., Kim, N.H. *et al.* (2015) An integrated system for precise genome modification in *Escherichia coli*. *PLoS One*, **10**, e0136963.
125 Chien, A.C., Hill, N.S., and Levin, P.A. (2012) Cell size control in bacteria. *Curr. Biol.*, **22**, R340–R349.
126 Kepes, F., Jester, B.C., Lepaqe, T., Rafiei, N. *et al.* (2012) The layout of a bacterial genome. *FEBS Lett.*, **586**, 2043–2048.

5

Functional Requirements in the Program and the Cell Chassis for Next-Generation Synthetic Biology

Antoine Danchin[1], *Agnieszka Sekowska*[1], *and Stanislas Noria*[2]

[1] *Institute of Cardiometabolism and Nutrition, 47 boulevard de l'Hôpital, Paris, 75013, France*
[2] *Fondation Fourmentin-Guilbert, 2 avenue du Pavé Neuf, Noisy le Grand, 93160, France*

A popular subject of hype, synthetic biology (SynBio), is described as a domain that will solve many of the catastrophic consequences of the human demographic explosion. Yet, the vision that stems from the engineering stance of this avatar of biology is seldom emphasized [1]. SynBio is uncommonly fruitful because taking life as an engineer would allow us to invert the classic view, where structure predates function, by placing function first [2, 3]. Innovation is a built-in consequence of engineering because it commonly originates from a top-down approach. It is based on functional analysis [4, 5], a methodology that endeavors to uncover, list, and organize the needed functions before they are implemented in the design of a particular contraption. This chapter illustrates the constructive role of engineering with a sample of SynBio-related functions relevant to the architecture of the genetic program connected with its associated host cell. Following trends developed by other investigators involved in the development of SynBio [6], we hope that introducing the logic of engineering will spur novel types of studies that will, eventually, result in successful applications of SynBio and more generally develop the future of biology with a fresh mind-set.

5.1 A Prerequisite to Synthetic Biology: An Engineering Definition of What Life Is

Engineering has tight relationships with what we recognize as science, created in Greece some three millennia ago. To see how it contributes to conceptual developments in our contemporary understanding of life, let us briefly recapitulate how engineering was associated with the history of science [7]. While inventing writing, our predecessors began to organize the world they live in by making inventories: herds of animals, bushels of grain, and stars in the heaven. The outcome of this effort had to be organized so as to retrieve and make the best use of the corresponding knowledge, when and where needed. Maps of the sky, of the

Synthetic Biology: Parts, Devices and Applications, First Edition. Edited by Christina Smolke.

land, and of the city were written down, drawn, and discussed. This led some to witness recurrent events within the records. Using repeated observations as marks led our forefathers to derive useful applications, in computation and in understanding and predicting how the world would fare. In parallel and in a reciprocal activity, making tools, buildings, and machines resulted from accumulated understanding of repeated features (e.g., in using metals, in particular, iron, at some point). In turn, this brought forth further understanding via all kinds of explorations, both in the concrete world and in abstraction. Engineering allowed those who ruled the city to tag events in time (including the regularity of days and seasons) and to measure the flow of time, in parallel with measuring positions in maps and lengths in space. The way we collect huge amounts of data today is not without similarity with the situation in this ancient era, making it, again, quite fit for the development of engineering.

Still, this activity was applied essentially to inanimate objects. Apart from the domestication of plants and animals, life mostly escaped the engineer's hands because it was so natural: it appeared everywhere, independent of man. Spontaneous generation was not the exception – it was the rule. You just had to let a broth stand in the air to see it losing its transparency and becoming full of worms. The consequence is that it took very long to think that life could also be open to engineering. We witness a follow-up of this attitude in today's reluctance of some to accept SynBio as the continuation of this prescientific attitude: after all, rational plant breeding has but two centuries of age (at most [8]), and we still witness sequels of the ancient idea that the moon directly influenced plant growth (see [9] for reference). Pasteur discovered that life was associated with dissymmetry. This led scientists to begin to see biology as a particular development of chemistry, at a time when the frontier between organic and inorganic matters had begun to vanish. *La dissymétrie, c'est la vie* (dissymmetry, this is life!) declared Pasteur. Indeed this claim pointed out the existence of some efficient and somehow easy selection process that would trap and carry over, within living organisms, some of the physical dissymmetry present in the universe. Pasteur remained a vitalist, but Justus Liebig and Claude Bernard, each one in his own way, propagated the idea that chemical processes were at the root of life. In short, their works asked for some definition of life that would be useful for an engineer wishing to (re)construct a living entity. It also, unobtrusively, pointed out a role for information, an overlooked currency of reality. It is high time today to put biology in the light of engineering.

The most successful engineering paradigm of cells and organisms is that they behave as machines running a program [10–14]. The basis for genetic engineering has been the development of techniques that allow investigators to synthesize pieces of genetic programs meant, oftentimes, to express genes into proteins of industrial interest [15]. In an early work, James Danielli saw that engineering could extend to the synthesis of life by combining individual bits and pieces into a functional entity [16, 17]. Despite this deconstruction/reconstruction procedure, it was long asserted that machine and program were intimately linked together and inseparable (see [18] for a justification of this negative view).

The onset of DNA-based technologies such as transfection of viral DNA into cells [19] and genetic engineering [15], associated with recognition that horizontal gene transfer made a considerable fraction of bacterial genomes [20], and finally whole genome transplantation [21] was a turning point. They established that the machine and the program are indeed separate entities, exactly as the operating system (OS), and the computer can be physically told from one another [10, 12]. It has been now possible to synthesize viral genomes in such a way that they comply with a man-made design [22, 23]. The statement found in rearguard discussions that the comparison between cells and computers is not valid because there is a considerable amount of information in the cell beside its genome cannot be retained as a final argument. Indeed, the situation is exactly the same in computers, human artifacts that fare well. Nobody would argue that the tablet or the PC do not carry a considerable body of information. The proof is that a CD carrying an OS is useless in the absence of the information carried by the machine that runs it. Yet nobody would argue against the fact that computers work, provided they can read a support carrying a matching OS.

Of course, this is not the whole story: besides program and machine (the "chassis" of SynBio specialists [24]), the cell, as the computer, needs to process energy, a feature that is not implemented in the abstract ancestor of the computer, the Turing machine. Furthermore, there is a need for construction and maintenance, which implies fluxes of matter, a currency of reality that is also absent from the purely informational Turing machine. In living organisms these essential functions are fulfilled by metabolism. Life can be witnessed only when metabolic fluxes can be measured, with "dormancy" labeling the limbo between life and death. In summary, life combines a *program*, a *machine* reading and expressing the program, and a *metabolism* managing matter and energy fluxes to run the program in the machine. Finally, a living organism works through an ultimate constraint: it must produce a progeny. Functions pertaining to that particular process make the core of the present chapter. Using functional analysis to understand the making of life, with emphasis on the processes just summarized, we propose here a set of developments that emphasize the mutual interaction between the program and the chassis, a setup essential to master for the future of SynBio.

5.2 Functional Analysis: Master Function and Helper Functions

The success of genome transplantation into recipient hosts – the founding experiment of next-generation SynBio [25] – is allowing scholars to look into biology with new eyes. To go further, we apply here the agenda of functional analysis [4, 5] to cells considered as "machines" or "automata," where a program can be explicitly told from the machine that runs it (Figure 5.1). When trying to understand how an organism can be fit for a particular niche, we first split its biological functions into two functional categories, at least one *master function* and associated *helper functions* meant to achieve the target of the master function [26].

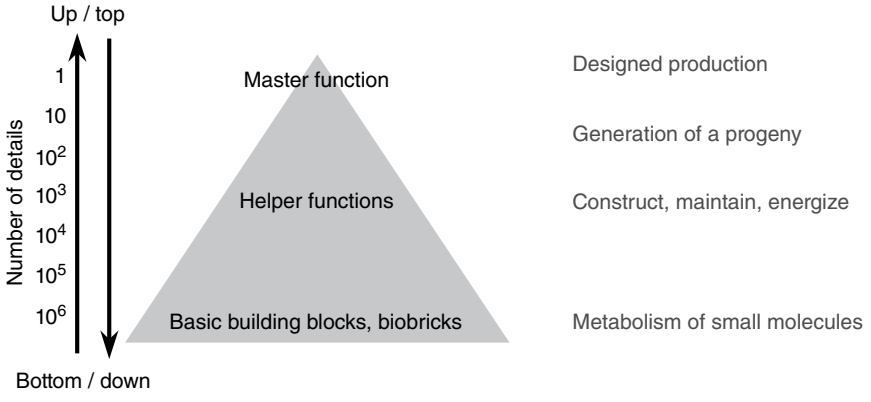


Figure 5.1 A schematic view of functional analysis [5]. Master and helper functions are as defined in the text.

Illustrated in a human artifact, the master function of a printer would be to print documents. The associated helper functions would be supplying paper, ink, electric power, and so on. Other helper functions would correspond to the design of the printer's chassis. When cells are envisioned as factories, their designed master function is production of some compounds. However this is entirely dependent on the ability of cells to multiply while replicating both their own program and the proper SynBio program construct, thus yoking the human construct to the cell's master function (multiplying).

While it is somewhat difficult to identify it without ambiguity, living organisms appear to display two intertwined master functions. The most obvious one is "to make a progeny." A myriad of helper functions have evolved to allow this master function to operate, and the huge variety of living organisms reflects this situation. This perspective (master function/helper functions split), however, remains fairly open. For most (this is the common view), "propagating life" is the destination of life. However, we must consider an alternate view, where "exploration" would be the master function, with "propagating life" as the immediately downstream helper function to that particular master function. Life would thus be a particular physicochemical process carrying further the intrinsic propensity for exploration carried over by all entities present in the universe (following the second law of thermodynamics that tells that physical systems will tend to occupy as many space and energy states as they can). Here, we favored the first choice, avoiding innovation – a major consequence of exploration – as a core property of SynBio constructs: who would like to fly in a plane that could modify its wings and engines in flight? We consider in what follows that the most general goal of SynBio is to make a reproducible automaton meant to produce compounds of preset design. This ranks exploration as a helper function that generally must be placed under command in SynBio constructs, and possibly even totally inactivated. We note however that our approach is operational and not directly linked to the concept of fitness that would entail a complementary discussion [27]. It is likely that the next decade will witness hot debates in this domain.

5.3 A Life-Specific Master Function: Building Up a Progeny

Life perpetuates itself. A sterile organism may still be alive, but it misses a key property of life in that it does not have a progeny. Indeed, its very existence is simply borrowing time: maintenance of a machine linked to a program, both doomed to age and die, can hardly allow long-term survival in an ever-changing environment (for a discussion, see [28] and references therein). Some animal societies have classes of sterile individuals, but they are always firmly connected to a fertile lineage. If life were only composed of infertile individuals, it would already be extinct, unless there existed a steady and speedy process of spontaneous generation with a creation time shorter than the life-span of individual organisms. This is more than unlikely and does not, anyhow, fit with the chemistry of life as we know it. We will therefore accept that life is tightly coupled to the making of a (young) progeny.

Considering this process, we can see that the ultimate destination of the genetic program is to make a copy of itself within a copy of the machine that runs the program. "Copy" here must be defined. How are the processes of program copying and that of cell copying linked together? Remarkably, the actual concrete copying process differs whether dealing with the program, or with the machine: the program is *replicated* in most of the cell's progeny (i.e., it makes *exact* copies of itself), while the machine's future is much sloppier, wherein it is only *reproduced* (i.e., it makes *similar* copies of itself) [27, 29]. To this dichotomy two time scales are associated: replication is trustworthy for many generations, while reproduction makes copies that vary rapidly over time. Genome transplantation experiments, such as those using synthetic genomes [25], give us a vivid illustration of this functional dichotomy. Extracted at the end of the experiment and sequenced, the synthetic genome of the bacteria in the recovered colonies is identical to that which has been transplanted in the host. By contrast, the machinery, and even the cell's shape, differs in the initial host and in the cells making the final colonies (Figure 5.2). In terms of engineering, this is somewhat unusual, although we all know of man-made devices that have been progressively modified, as was Theseus boat (that did not keep a single original of its boards after some time [11]). The parent machine has aged, and its components have been replaced by new ones. In the transplantation experiment, this regeneration process required the use of a new program, differing from the parent one that had been destroyed, thanks to an astute genetic design [21]. As a consequence, during multiplication, the program that was used is that of the transplanted genome, directing the synthesis of entities that differ from those of the initial host machine.

This state of affairs is far more general than that in the transplantation experiment: as in any life form, the components of any SynBio construct age and are replaced; in parallel, the environment changes and some components are no longer required and are diluted out while others are expressed. In short, while the program may remain the same, the machine that runs it is quite variable. It keeps however its main functional (abstract) properties: reading and expressing the program, and directing the construction of a progeny, while monitoring the state of the environment, extracting proper resources and discarding useless or

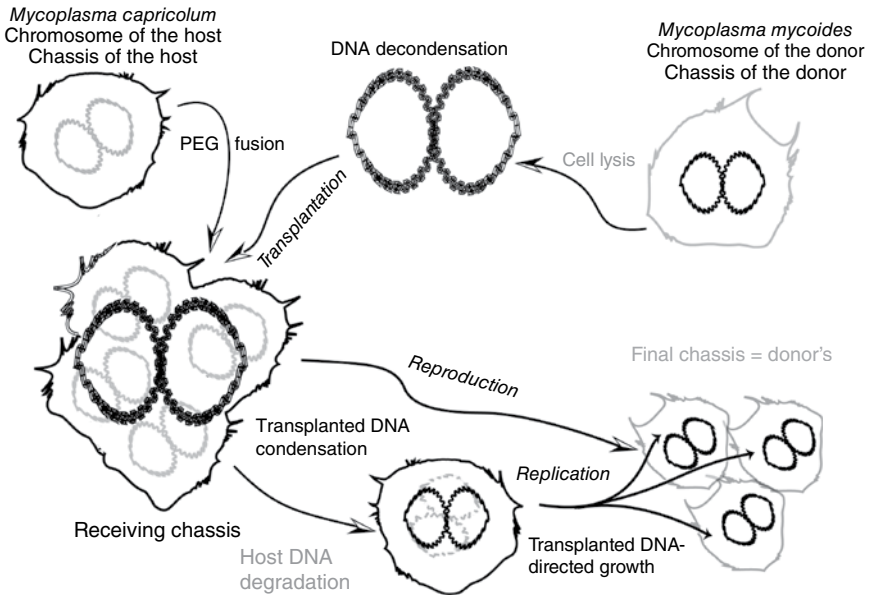


Figure 5.2 Replication of the program, reproduction of the chassis. The sequence of the *Mycoplasma mycoides* DNA transplanted into *Mycoplasma capricolum* is identical to that at the end of the experiment. Transplantation has triggered degradation of the *M. capricolum* DNA, while the DNA of *M. mycoides* replicates and dilutes out the component of the initial host. At the end of the experiment, the components of the cells are identical to those of *M. mycoides*, not to those of the recipient *M. capricolum*.

worn-out components. The relationship between the machine and the program is central to this essential interaction. This situation is also common in contemporary computers, which remember our past actions and do not behave today as they did some time ago, improving their adaptation to our wishes as time elapses. In cells, this corresponds to exploiting an information that is not directly present in the genes, but, rather, to a contextual information present in the way genes are placed (and sometimes tagged by specific biochemical processes) in the genome and its disposition within the cell as well as in the ultimate matter making the genome. We note in passing that the transplantation founding experiments tells us something more in terms of functions. It uncovers the first hidden functional constraint on the genome structure: the chromosome needs to be compacted to fit a small volume, and this is why (Figure 5.2) the transplantation experiment requires as a first step the making of a syncytium to accommodate a decondensed DNA molecule [27].

5.4 Helper Functions

For life to keep going and to develop into a descent, a chore of helper functions is needed. These functions operate at different levels. They are organized along hierarchies that are segmented (like organs in an animal body) or branched

(like trunk, branches, and leaves of a tree). To take this hierarchical view into account, the functional part of the gene ontology effort has endeavored to organize functional data as well as structural data [30]. The outcome of this remarkable effort has still to be considerably improved, and it is likely that much innovation will appear there in the near future [31].

Just to name a few helper functions, we find excerpts from an unlimited list: a way to go forward and uncover unexpected functions would be to use the list of all the verbs present in a particular language:

Making a progeny, with associated functions:
Construction of biomass
Replication of the program
Division (separating the progeny from the parent organism)
Maintenance
Making the progeny young, that is, separating between young and aged entities

Exploration is the function that could also be considered as a master function for all living organisms, as life is doomed to explore its environment. It implies either harnessing movements of the environment (spores or seeds propagated by wind), constructing appendices allowing the organism to move (flagella in microbes, limbs in animals), or harnessing features of the environment to movements (light with phycobilisomes, magnetic field with magnetosomes, etc.).

Each one of these functions is achieved using a lower level of helper functions, some of them universally required for replication and reproduction, while others are used by the organism for moving and occupying a particular niche [32]:

Transport (in and out): Extraction of chemical compounds from the environment, and getting rid of waste
Circulation
Sensing
Management of energy
Storing
Shaping and maintenance of the cell structures
Degradation/resynthesis
Protection

To make this analysis explicit and concrete, we explore now some of the topics that are central to place the genome in the cell's context. Let us split some of these helper functions along the dominating contribution of each of the five universal currencies of reality: matter, energy, space, time, and information, with emphasis on constraints on the genome (including its assembly).

5.4.1 Matter: Building Blocks and Structures (with Emphasis on DNA)

Formation of a cell begins with metabolism. For decades, this topic was considered as a boring haphazard collection of chemicals. It was taught in classes where students tried to remember by heart the lists of compounds, not really understood as following much logic. Yet, there is a clear logic of metabolism that begins to be deciphered [33]. This logic is consistent in terms of the physics of matter.

However, the local stability and relevance of retained pathways is not optimized in terms of what would be a human engineering design. This implies that optimization – in terms of sparing energy and matter – will be at the core of next-decade metabolic engineering. To describe this logic would ask for a whole textbook, and we will just enumerate here some of the rules that are beginning to emerge. As the material support of the genetic program, DNA synthesis from nucleotides required for replication of the genome will be described in some details.

The atoms of life are not random: carbon, hydrogen, nitrogen, and oxygen compose living organisms because they are prone to combine via chains of covalent bonds that are stable at the temperature of the Earth's surface; heavier elements would not retain this property in general. Sulfur (together with iron) is added to the list because of constraints well understood in scenarios of the origin of life [34]. This atom is also a remarkably versatile support for electron transfers. It exists in biological compounds in redox states going from –2 to +6, and this useful property made that it has been retained in the course of evolution [35]. Phosphorus is unique when combined to oxygen, as phosphate bonds are prone to hydrolyze (hence easy to disrupt in water), yet metastable (hydrolysis generally requires a large activation energy). This property, which allows phosphate compounds to store energy, is the reason phosphorus belongs to the core atoms of life on Earth [36]. This constraint is essential to remember when looking for xenologous BioBricks meant to construct genetic programs for xenobiology. The role of phosphates provides us with a strong argument in favor of the engineering stance. Had the way engineers think be favored, arsenic would never have been considered as a substitute for phosphorus [37].

Phosphorus is a core component of nucleic acids, enabling a specific metabolic driving force associated with hydrolysis of pyrophosphate (polymerization of nucleotides is reversible; therefore going forward requires an irreversible step). Furthermore, the organization of phosphate metabolism drives the nucleotide composition of the genome in a way that is not still completely understood. Indeed, deoxyribonucleotides are essentially synthesized from the ribonucleoside *di*phosphates, not *tri*phosphates. This constraint is likely derived from the selection pressure that uses a metabolism developed in the three-dimensional environment, for synthesis of a linear molecule. This has a remarkable consequence for pyrimidines, as their anabolic pathway produces uridine diphosphate (UDP), but not cytidine diphosphate (CDP). This should lead to deoxyuridine diphosphate (dUDP) and then deoxyuridine triphosphate (dUTP), while input of U in DNA must be avoided at all costs via a complex set of pathways. Missing CDP would require an indirect process to make deoxycytidine diphosphate (dCDP) and then deoxycytidine triphosphate (dCTP) [38] (Figure 5.3). The consequence of this imbalance is that, in most cases, the genetic program tends to be progressively enriched in A+T nucleotides [39, 40]. The degradosome (with its exosome counterpart in Eukarya) is the machinery that resolves this hurdle. It allows buffering and equilibration of nucleic acids composition via degradation of RNAs by phosphorolysis (directly producing the much wanted nucleoside diphosphates (NDPs), in particular CDP, that are the precursors required for DNA replication) while coupling the fluxes of nucleotides with energy resources [38, 39, 41]. Furthermore, the physical relationship between phosphate

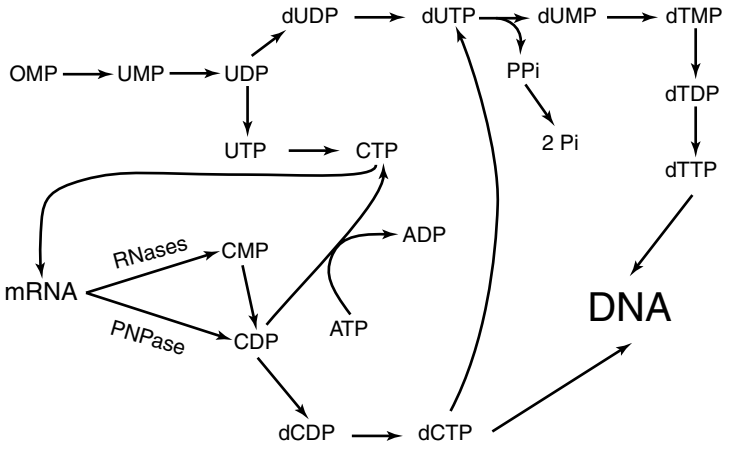


Figure 5.3 Excerpt of the metabolism of pyrimidines and DNA synthesis. The building blocks for DNA stem from NDPs, not nucleoside triphosphates (NTPs). This creates an imbalance in the case of cytosine, because CDP is not produced during the *de novo* synthesis. This explains why, in general, C is the limiting nucleotide, driving A+T enrichment of the genome in most situations. Nucleoside diphosphokinase is reversible; however ATP is in excess over adenosine diphosphate (ADP), so that production of CDP is limiting via this route. CDP comes mainly from mRNA turnover via phosphorolysis (polynucleotide phosphorylase) or RNase activity, with further phosphorylation using cytidylate kinase.

metabolism, replication, and transcription is likely to have considerable bearing on the genome organization within the cell (it is at the root of the preservation of a nucleus in eukaryotes).

The genome backbone phosphate is not strictly universal. Some organisms use a variant where the usual phosphate group is phosphorothiolated [42]. This modification, which can be used for specific recognition/folding processes and provides the cells with a protection against oxidative stress [43], is a first hint that SynBio could evolve toward xenobiology (i.e., the use of nonstandard building blocks for the construction of synthetic cells [44]). A further indication of this possibility is the presence of diaminopurine instead of adenine is some cyanophages [45]. Finally, de Crécy-Lagard and coworkers showed that a 7-deazapurine derivative can replace guanine in functional DNA [46]. Knowing that DNA methylation can be used to control gene expression [47], the idea that other modifications may have a similar role is straightforward. A track for the future analysis of the distribution of phosphorothiolated sites or input of 7-deazaguanines has not yet been undertaken, and their role in gene expression is not known. In general, there is still considerable room for exploring the presence and role of nucleic acid modifications [48].

Amino acids make the primary sequence of proteins, while many more exist in metabolic pathways (be it only as the result of catabolism of posttranslationally modified proteins). Proteinogenic amino acids are far from random, however, as several are fairly easy to synthesize (the smallest ones) and can be converted into one another at low energy cost [49], while they are split into three major physicochemical properties highly relevant to water as a universal albeit physically unusual

solvent, essential for the development of life as we know it [50]. Proteinogenic amino acids are hydrophilic, amphiphilic, or hydrophobic. Surprisingly, proline is not an amino acid but an *imino* acid, and this has considerable consequences for translation, with requirement of a specific elongation factor [51]. As building blocks of proteins, these molecules need to be activated as aminoacyl adenylates and loaded onto the 3'OH extremity of the ribose of an acceptor tRNA molecule. This process introduces considerable constraints in the selection of amino acids relevant to translation: for example, ornithine, homoserine, and homocysteine will cyclize during the process and create toxic dead-end compounds [35, 52]. Norleucine and selenomethionine can substitute for methionine [53], and this changes activity only in a restricted number of proteins but is deleterious when methionine is limiting as the methionine side chain is specifically used in some metalloproteins, for example [54]. 2-Aminobutyrate is an analog of cysteine, and its concentration must be stringently controlled as it mimics cysteine metabolism. Furthermore, the presence of non-proteinogenic amino acids in cells implies that they are prone to affect negatively translation accuracy via their wrong incorporation into proteins. Hence the cell must cope with this hurdle either by maintaining a very low level of non-proteinogenic amino acids or by modifying them (generally by *N*-acylation and sometimes *N*-methylation) so that they do not enter the wrong pathway [55].

Interestingly, this ubiquitous protection pathway (in the sense given by engineers in organic chemistry) is likely to have been recruited for other helper functions such as further protection or regulation. For example, ribosomal proteins are generally acylated, but the exact function of the modifications remains unknown, except for an obvious coupling with metabolism and a protective role for reactive amines [56, 57]. In the case of nucleotides, it may well be that formation of the triphosphates, besides providing a way to drive forward biosyntheses via pyrophosphate hydrolysis, plays also the role of a recognition group, resulting in the selection of a subset of nucleotides for insertion in polynucleotides.

Another mode of metabolism organization derives from a distinctive match between matter and space constraints. Carbon chemistry allows formation of a considerable number of specific stereoisomers (remember Pasteur's exclamation) that are recognized by enzyme cavities in a highly space-constrained way. Proteinogenic amino acids are of the L-type. As a consequence most proteases and peptidases are active on chains made of these stereoisomers. This opened up the possibility of a selection pressure, leading to protease-insensitive protective structures that evolved toward containing the D-isomers (e.g., antibiotics [58]). Another most important selection pressure is on compounds that are similar to amino acids, with a hydroxyl group in the place of the alpha-amino group, making them good mimics of amino acids. For example, glycerate is quite similar to serine and could take its place in many enzymes, an unwanted stereochemical toxic property. D-Glycerate is therefore the preferred stereoisomer [59]. This has consequences on the make-up of nucleic acids. Because of the link between the latter metabolite and those involved in glycolysis/gluconeogenesis, this stereochemical constraint explains why most biologically relevant carbohydrates, ribose, and deoxyribose, in particular, are of the D-isomer type [60].

This metabolic constraint must be taken into account when planning to derive novel nucleic acid analogs for next-generation SynBio.

At a more integrated level of the hierarchical organization of life, multicellular organisms have developed an extraordinary diversity of macromolecular materials that work as frames, protectants, buffers, motors, signals, traps, and so on. DNA itself is known to belong to the structural polyanionic polymers, as it is, for example, a component of biofilms [61], which introduces a fitness property that has nothing to do with its coding capacity. We have seen that there is anyway a significant selection pressure to increase its length in order to match its synthesis with that of the bulk of the cell (see Section 5.4.3). Exploration of chemical diversity both in terms of small metabolites (see, e.g., [62–64]) and in terms of macromolecules is expected to develop considerably (see, e.g., [65–67]) in the next decade. The corresponding genetic program implementation within the cell will need to be explored in depth. Here again thinking as an engineer will come as an asset for innovation.

The list of engineering constraints on the matter used in living organisms is unlimited. The examples presented earlier are just meant to illustrate the way we should presently consider metabolism. In another dimension, metabolites of industrial interest, such as isobutene, are and will be produced by reprogramming and setting up synthetic pathways [68]. This will entail production of molecules that may react with components of existing cell components, including DNA. To end up with high yields, metabolic engineering will need a deep reflection on metabolite reactivity within the confined medium of the cell, a topic that has mostly been restricted to the study of reactive oxygen species [69]. In particular it seems obvious that the chromosome must be protected, as much as possible, against reactive metabolic intermediates (we saw previously that phosphorothioation is a solution uncovered during evolution). Management of waste will also be a major topic to be developed (be it only to limit carbon dioxide production).

5.4.2 Energy

Management of energy is central to life. It has long been established to be associated with electron and proton transfers and with storage as energy-rich phosphate bonds. The motto "better lose energy than control" seems to dominate life. Much is known about the energy-related processes, but much also remains to be understood in terms of optimization. For example, despite their ubiquity (all cells contain these compounds), the role of polyphosphates has seldom been considered, despite their importance for energy management, regulation, and storage [70–72]. It seems likely that their contribution as the ultimate energy source (polyphosphates are minerals, hence particularly resistant to desiccation, radiations, and harsh environments) needs to be reconsidered, especially in terms of synthesis and usage during transition states, aging, and stresses [28]. Nucleic acids are also energy-rich molecules. This implies that some regions of DNA might in fact have a role as energy stores, besides their expected role in space management, gene coding, and regulation. In this respect, some organisms have a genome of huge size [73], without any apparent direct link with its coding capacity.

Novel features of electron transfers are related to the way protons and electrons can be transported within and between cells. A considerable effort has yet to be devoted to the production of hydrogen as an energy store [74–76]. Research on microbial fuel cells is expected to develop dramatically [77–79]. It has now been recognized that cells can make wires that conduct electricity, creating an entirely new field for management and extraction of energy via living systems [80–82]. Some cells make large syncytia, which requires management of the genome DNA and energy sources (polyphosphate in particular) in a way that is not yet understood [71]. The role of membranes is essential in building up and maintaining the electrochemical potential of the cell via vectorial transport. In the same way energy is stored in a variety of polymer compounds such as lipid droplets, carbohydrate polymers, polyphosphates, and so on. This introduces a specific link between energy and space, the role of which we now discuss.

5.4.3 Managing Space

In the cell, the three dimensions of space play together in a concerted fashion. The genetic program is stored by a molecule of DNA that can be considered as linear in the way it maintains its coding capacity; membranes organize space in two dimensions; finally the interior of the cell is three-dimensional. In terms of biosyntheses this has consequences that are seldom taken into account. Filling up the cytoplasm with proteins as the cell grows requires an increase as the cube of the cell's size (if the cell is spherical, less when it is of another shape) while placing proteins in the membrane would go as the square of the cell's size. This discrepancy introduces a considerable constraint on the length of the genome. It cannot be too short, which implies that despite a selective tendency to streamline the genome sequence because of the cost to maintain functional genes, there is an opposite tendency to fill it in with extra DNA sequences. Amplification of insertion sequences or similar structures and horizontal gene transfer can compensate for deletions. Overall insertions and deletions create an equilibrium that results in an optimum length, where the DNA length is considerably longer than that of the cell. Indeed, 4′,6-diamidino-2-phenylindole (DAPI) staining shows that the genome in itself occupies a significant proportion of the cell's volume [83], shaping it more like a three-dimensional structure, via folding into a Peano curve-like space-filling setup [84]. This constraint is likely to be important in the gene flow that maintains a particular genome length [85].

Chromosome DNA folds can be classified into three categories [86]: short range, of up to 16 kb (fitting with the local bias in codon usage [87]); medium range, over 100–125 kb (fitting with old observations of supercoiled DNA loops upon mild cell lysis [88]); and long range, over 600–800 kb (fitting with the size of the shortest bacterial chromosomes and associated with macrodomains [89, 90]). In Eukarya, the problem of the various space scales has been solved by the preservation of a nucleus accommodating the genome in a space much smaller in general than the size of the cell and multiplying membranes (in particular the endoplasmic reticulum) to couple protein synthesis with occupation of the cytoplasmic space. Chromosome folding requirements appear to impose sequence constraints that create ubiquitous 11 bp periodic patterns, the "class A flexible

patterns" (previously identified as their core sequence, ApA dinucleotides [91, 92]), spanning 5–10 helix turns, and present every few kilobases of the whole genome. These patterns have been repeatedly observed and usually thought to result from the local enrichment of ApA dinucleotides, but likely to result from more complex patterns [93–95]. This constraint is so strong that it appears to bias the nature up to one in five nucleotides in the genome [95]. In general A-tracts have been related to DNA curvature, and they are expected to play a considerable role in DNA compaction and regulation of gene expression [96–98]. The importance of this feature has not yet been explored in SynBio constructs.

Managing space is also essential to organize gene expression (references in [84, 90, 99, 100]). Indeed, while the DNA molecule is a linear structure, the membrane is a 2D structure and the cytoplasm is a 3D structure; they all need to work in concert. Allowing coordination of the different space scales, gene expression [101] and distribution of genes within transcription units are finely tuned in most Bacteria and Archaea, in particular in terms of coordination of metabolic fluxes [84, 102]. For example, in the lactose operon, the gene for cytoplasmic beta-galactosidase, *lacZ*, is separated from that of the membrane protein lactose permease, *lacY*, by a regulatory transcription attenuator. This results in considerably less expression of the distal genes *lacY* and *lacA*, as compared with that of *lacZ*, and allows matching the production level of the cytoplasmic enzyme with that of the membrane transport protein [103]. In general there is a relationship between the genome organization and the pattern of transcripts and protein distribution in the cell [86, 104].

The genome DNA is considerably longer than that of the cell, and this allows folding of the chromosome in a way that can compensate for the one dimension/three dimensions dichotomy. Furthermore there seems to exist a relationship between the overall cell architecture and that of the genome; Tamames and coworkers found a remarkable correlation between the distribution of genes in the *mur-fts* gene clusters and the overall shape of the cell [105, 106]. This observation may fit with the view that transcripts are systematically distributed in specific regions of the cell, as shown by local biases in codon usage, forming islands 10–30 kb long [87] in agreement with the data reviewed by Willenbrock and Ussery [100]. In general, analysis of the folding of the chromosome revealed the existence of a core structure linking together between 12 and 80 loops per chromosome [88, 107]. Many studies have explored the role of the distribution of the genes in the bacterial chromosome, in particular with the prospect of improving gene expression in biotechnological constructs (see [108] for further references). Despite the widespread view of the chromosome as extremely plastic, it rapidly appeared that while some regions were prone to harbor a variety of genes, others remained fairly constant. Indeed, macrodomains organization appears to display rigid constraints that limit genome plasticity [109]. This was further illustrated with the comparison between a large number of *Escherichia coli* strains [110, 111]. It was also found that functionally related genes clustered together into islands in a way that should have considerable impact on gene expression [100, 108, 112].

The two extremes of gene distribution are clustering and its opposite, uniform distribution (which creates an apparently periodical distribution, so that

noticing a period should not be taken as particularly significant). Steady random insertion of genes via horizontal gene transfer will go toward creating a uniform distribution of the genes that are most important for the cell life, while frequent deletion will tend to make them cluster together [85]. Another well-identified constraint in the genome of fast-growing bacteria results from the fact that genes located near the origin of replication will tend to be in higher copy number in the growing cell as compared with genes located near the terminus of replication [90]. This difference is also reflected in the distribution of codon biases classes [87]. When long enough, the bacterial chromosome is further organized into macrodomains that are insulated from one another and are essential for genome packaging [89, 113–115]. The presence of plasmids or several chromosomes alters this distribution [116]. Finally, there is a significant pressure for important genes to be transcribed from the leading replication strand in order to avoid transcription/replication conflicts [117]. Knowledge of these organization constraints is essential for optimizing gene placing in SynBio constructs.

Management of space is further associated with several kinds of functional structures, exoskeleton and endoskeleton, scaffolds, and contractile proteins such as actins and myosins in the cytoplasm. How do the corresponding macromolecules know where and when to go as the cell grows, changes its shape and eventually divides? In this context, it was revealing to discover that Bacteria and Archaea were not different from Eukarya, having a variety of structuring proteins, often associated with the inner membrane and contributing to the overall shape and functional properties of the cell [118]. As a common feature, the prokaryotic and eukaryotic cytoskeleton proteins couple energy requirement, via adenosine triphosphate (ATP) and/or guanosine triphosphate (GTP) utilization in active (energy-requiring) mechanisms to effect structuring functions and manage movements. The corresponding logic of engineering design has yet to be uncovered. A family of proteins, the structural maintenance of chromosome (SMC) proteins, manages chromosome spatial arrangement and replication, at the expense of energy [119]. As another example of versatile functional design, membrane protein topology is coupled to functional addressing, with recently recognized proteins with dual topology [120]. Interestingly, there is a coupling between genome evolution and these proteins: genes in families containing dual-topology candidates occur in genomes either as pairs or as singletons, and gene pairs encode two oppositely oriented proteins whereas singletons encode dual-topology candidates [121].

Finally, getting in and out of the cell is essential: the cell has to manage the influx of compounds used to construct biomass and create energy. It has also to dispose of waste. These processes occur at the membrane, using a variety of structures. Often, the cell has to extract useful compounds from an environment where they are considerably diluted. This requires an energy-dependent active transport that concentrates molecules up to a thousand-fold or more. This essential engineering process has a trade-off: if the outside concentration of the compound increases suddenly, the influx will build up an unbearable osmotic pressure that will require coupling modification of the influx molecule and safety valves in order to prevent the breakup of the membrane [122, 123].

Another question that must be answered is the way the influx of protons is distributed within the cell. As the membrane-associated rotor of ATP synthase or the flagella motor leaks in protons at a fast rate, their influx must be coupled to a steady average amount of "free" protons in the cell that is extremely low (typically, if there were such a pH as 7.6 in an *E. coli* cell, this would mean about 15 free protons per cell at any time). The way protons are disposed of so that on average such a small number remains free is an entirely open question that requires understanding the way water is organized in the extremely crowded environment of the cytoplasm. This situation has considerable consequences in particular for highly charged molecules such as nucleic acids. This is not yet really understood [124, 125]. An alternative to safety valves is storage by polymerization, a function fulfilled by a variety of structures and compartments [126], and polymerization of nucleotides is a way, rarely considered, to buffer osmotic pressure.

5.4.4 Time

The idea that time and transitions are essential in shaping molecules and organizing cells is also central to the understanding of the addressing, organization and motion of proteins within the cell and its membrane. The role of time will be one of the most important features of the development of SynBio in the next decade. This is because in most research, studies of evolution and phylogeny aside, there has been a tendency to account for life in synchronous terms. For example, the recent descriptions of the way DNA is folded in cells provide us with a fairly static view [127–129]. Yet, it is obvious that except in dormant states, DNA is highly flexible and mobile, with movements triggered by transcription and all related processes that maintain supercoiling, as opening up the double helix locally will trigger a deformation that will propagate [127, 130, 131]. It is likely that the organization into macrodomains is fit to coordinate gene expression [113], including when transcription involves time-dependent movements of the DNA template.

Considering cells and organisms as computers, making computers exposes a considerable possible limitation, where time plays a central role, resulting from the fact that expression of the genetic program is highly parallel. Parallelism implies that a variety of clocks allow synchronization of gene expression processes [27]. This need for synchronization is likely to be another constraint that organizes the genome into macrodomains [89]. Indeed, clocks are found everywhere in life: coupling of gene expression with seasons [131], circadian rhythms [132], and many other kinds of clocks, unrelated to obvious environmental parameters [133]. It has been known since the nineteenth century that circuits with relevant feedback loops could end up with oscillating properties, de facto creating clocks. It is therefore quite trivial to find clocks based on regulatory gene expression circuits, an expected property that nevertheless became quite fashionable several decades ago, hiding more interesting roles of time. By contrast and more interestingly, other intrinsic clocks, for example, based on the aging half-life of macromolecules (such as resulting from isomerization of asparagine and aspartate in proteins [27, 134]) may bring about unexpected uses of

time. In general, the importance of time has been underestimated, in particular because laboratory conditions are most often meant to provide steady-state invariable conditions. Time-dependent pattern formation, a basis for multicellular body plan, must be explored with novel approaches [135]. Finally, the role of ubiquitous transitions (shifts in temperature, light, metabolites supply, interactions with other organisms, and simply aging) will certainly need to be explored much more in-depth for large-scale SynBio applications. The time scales of DNA movements have not been explored in-depth, and, if relevant, this missing knowledge might become a limitation for the future of SynBio constructs.

5.4.5 Information

SynBio uses cell factories that associate a program with a chassis. As previously discussed, the transplantation experiment that implemented a program that did not match the receiving host chassis [21] demonstrates the physical material separability between machine and program [10, 12]. It also emphasizes another point, where information is central: while, at the end of the experiment, the donor's program is identical to that at the beginning, the final machine (*Mycoplasma capricolum*) differs from the initial host machine (*Mycoplasma mycoides*) (Figure 5.2). This implies that some specific input of contextual information (gene expression in a particular environment, at a particular time), and not directly related to the information carried over by the genetic program, has been involved. In the same way, construction of a young progeny from aged cells demonstrates that there is a specific management of information by cells, in a way that is highly reminiscent of the way Maxwell's demons operate [27, 136]. Briefly, creating a link between information and entropy, Maxwell introduced the idea of a hypothetical being, later seen as a "demon" that uses an in-built information-processing ability to reduce the entropy of a homogeneous gas (at a given temperature). The demon is able to measure the speed of gas molecules and open or close a door between two compartments as a function of the molecules' speed, keeping them on one side if fast and on the other side if slow. This behavior will build up two compartments, one hot and one cold, reversing time and acting apparently against the second principle of thermodynamics. In the same way, proteins such as septins prevent aged proteins to go from the mother cell to the daughter cells [137] or organize cell division [138], using energy to reset their state to ground level [27, 136].

Information is split into several components: a genetic memory, carried over by DNA via faithful replication, epigenetic memory that reproduces a particular state of the chassis, including a specific organization of gene expression, and a variety of processes managing information transfers. DNA replication uses an asymmetrical nanomachine that breaks the DNA double helix opened at a specified origin and starts elongating a continuous strand in the 5′ to 3′ direction. The process is straightforward in replication of the leading DNA strand. By contrast, replication of the lagging strand poses major structural problems. Indeed, replication of that strand requires a considerable length of single-stranded DNA that must be protected by specific complexes; it also requires management of multiple initiation complexes, in contrast to replication of the leading strand, which

may start from a unique replication initiation locus [139]. This dissymmetry implies that the error replication rates differ on each strand, with different proofreading systems. Many proofreading processes exist, including those, such as the ATP-powered RecBCD nanomachine that takes care of double-strand breaks in *E. coli* [140].

Transcription operates with constraints similar to those of leading strand replication. Following transcription, the protein biosynthetic machinery brings together complexes composed of ribosomes, chaperones, and localization factors into similar actions (begin, elongate, and end). It also interacts directly with factors dedicated to disposal of protein fragments (generated during mistranslation, translation interruption, or premature termination) and more generally to protein degradation [141]. The genetic code accommodates 20 amino acids plus two variable ones, selenocysteine (coded for by UGA) and pyrrolysine (coded for by UAG). Remarkably it seems that in some organisms, the genetic code can be modulated via specific growth conditions [142] and that the UGA codon can be reassigned to a particular amino acid, differing from tryptophan or selenocysteine [143]. This implies that the genome could be read at levels of information much more elaborate than those understood until now. Nothing is known about the corresponding gene organization in the genome, but this opens up considerably the possibilities of information management in SynBio constructs.

Many other functions must be considered in the making of macromolecules and eventually implemented in SynBio constructs. Most deal with the fact that the threadwire machinery that makes macromolecules cannot fold them readily into their final proper three-dimensional shape (discussed in [27] to account for the hard time witnessed to succeed in genome transplantation) as well as in maintenance of the designed shape.

Finally, regulation is another key informational process. It is the main subject of most present SynBio experiments, many "BioBricks" being DNA segments used to construct regulatory logical gates, with strong emphasis on similarity with electronic circuits [144]. Some regulatory functions linked to sensing are regulated by the widely spread sensor-regulator two-component systems [145], where the channeling of information (separating channels is a challenge) has not yet been explored. Mechanical sensing is also important during cell growth, as well as when gases witness pressure changes [146]. Among the functions of information transfer, the control of metabolic and development processes is essential. Indeed, regulation lies at the core of the SynBio activities centered on the genetic program, and the bulk of the work dealing with BioBricks and the like aim at constructing sophisticated regulatory devices [147, 148]. This will not be explored further here as regulation is the focus of the vast majority of SynBio-devoted work [149].

5.5 Conclusion

SynBio rests on the description of living organisms as separating a genetic program from the machine that runs it. In general it is implicitly assumed that it is possible to use extant organisms as reference chassis into which on may

transplant artificial genetic constructs, with outputs that work well. Indeed, the proof of concept of this view has been repeatedly established, in constructing all kinds of circuits or metabolic pathways, showing that the key idea of the cell as a computer is at least conceptually viable. However, as industrial processes require both stability in time and high production, it is important that the proof of concept is followed by scaling up in making economically viable constructs. We have delineated here some of the constraints that must be taken into account to allow a smooth transition from the academic laboratory to the industrial scale.

Acknowledgments

Stanislas Noria is a network supported by the Fondation Fourmentin-Guilbert. The authors declare that they have no conflict of interest involved in this work.

References

1 Potthast, T. (2009) Paradigm shifts versus fashion shifts? Systems and synthetic biology as new epistemic entities in understanding and making 'life'. *EMBO Rep.*, **10** (Suppl. 1), S42–S45.
2 Danchin, A. (1999) From protein sequence to function. *Curr. Opin. Struct. Biol.*, **9**, 363–367.
3 Galperin, M.Y., Walker, D.R., and Koonin, E.V. (1998) Analogous enzymes: independent inventions in enzyme evolution. *Genome Res.*, **8**, 779–790.
4 Cole, E.L. Jr. (1998) Functional analysis: a system conceptual design tool [and application to ATC system]. *IEEE Trans. Aerosp. Electron. Syst.*, **34**, 354–365.
5 Fantoni, G., Apreda, R., and Bonaccorsi, A. (2009) A theory of the constituent elements of functions, in *Proceedings of the 17th International Conference on Engineering Design (ICED'09)* (eds M. Norell Bergendahl, M. Grimheden, L. Leifer, P. Skogstad, *et al.*), Design Society, Stanford University, San Francisco, CA, pp. 179–190.
6 Endy, D. (2005) Foundations for engineering biology. *Nature*, **438**, 449–453.
7 Dyson, F.J. (2012) History of science. Is science mostly driven by ideas or by tools? *Science*, **338**, 1426–1427.
8 Frängsmyr, T., Heilbron, J.L., and Rider, R.E. (eds) (1990) *The Quantifying Spirit in the 18th Century*, University of California Press, Berkeley, Los Angeles, Oxford.
9 Beeson, C.F. (1946) The moon and plant growth. *Nature*, **158**, 572.
10 Brenner, S. (2012) Turing centenary: life's code script. *Nature*, **482**, 461.
11 Danchin, A. (2003) *The Delphic Boat. What Genomes Tell us*, Harvard University Press, Cambridge, MA.
12 Danchin, A. (2009) Bacteria as computers making computers. *FEMS Microbiol. Rev.*, **33**, 3–26.
13 Quastler, H. (1953) *Essays on the Use of Information Theory in Biology*, University of Illinois Press, Urbana.

14 Yockey, H.P. (1992) *Information Theory and Molecular Biology*, Cambridge University Press, Cambridge.
15 Cohen, S.N., Chang, A.C., Boyer, H.W., and Helling, R.B. (1973) Construction of biologically functional bacterial plasmids in vitro. *Proc. Natl. Acad. Sci. U.S.A.*, **70**, 3240–3244.
16 Danielli, J.F. (1972) Context and future of cell synthesis. *N. Y. State J. Med.*, **72**, 2814–2815.
17 Danielli, J.F. (1974) Genetic engineering and life synthesis: an introduction to the review by R. Widdus and C. Ault. *Int. Rev. Cytol.*, **38**, 1–5.
18 Longo, G. (2009) Critique of computational reason in the natural sciences, in *Fundamental Concepts in Computer Science* (eds E. Gelenbe and J.-P. Kahane), Imperial College Press/World Scientific, London, pp. 43–70.
19 Epstein, H.T. (1968) Factors affecting bacterial competence for transfection and transfection enhancement. *Bacteriol. Rev.*, **32**, 313–319.
20 Medigue, C., Rouxel, T., Vigier, P., Henaut, A. *et al.* (1991a) Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J. Mol. Biol.*, **222**, 851–856.
21 Lartigue, C., Glass, J.I., Alperovich, N., Pieper, R. *et al.* (2007) Genome transplantation in bacteria: changing one species to another. *Science*, **317**, 632–638.
22 Chan, L.Y., Kosuri, S., and Endy, D. (2005) Refactoring bacteriophage T7. *Mol. Syst. Biol.*, **1** (2005), 0018.
23 Jaschke, P.R., Lieberman, E.K., Rodriguez, J., Sierra, A. *et al.* (2012) A fully decompressed synthetic bacteriophage øX174 genome assembled and archived in yeast. *Virology*, **434**, 278–284.
24 de Lorenzo, V. (2011) Beware of metaphors: chasses and orthogonality in synthetic biology. *Bioeng. Bugs*, **2**, 3–7.
25 Hutchison, C.A. 3rd, Chuang, R.Y., Noskov, V.N., Assad-Garcia, N. *et al.* (2016) Design and synthesis of a minimal bacterial genome. *Science*, **351**, aad6253.
26 Danchin, A. and Sekowska, A. (2013) Constraints in the design of the synthetic bacterial chassis. *Methods Microbiol.*, **40**, 39–68.
27 Danchin, A. (2012) Scaling up synthetic biology: do not forget the chassis. *FEBS Lett.*, **586**, 2129–2137.
28 Danchin, A. (2009) Natural selection and immortality. *Biogerontology*, **10**, 503–516.
29 Dyson, F.J. (1985) *Origins of Life*, Cambridge University Press, Cambridge.
30 Blake, J.A. *et al.* (2015) Gene ontology consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
31 Ruch, P. (2017) Text mining to support gene ontology curation and vice versa. *Methods Mol. Biol.*, **1446**, 69–84.
32 Acevedo-Rocha, C.G., Fang, G., Schmidt, M., Ussery, D.W. *et al.* (2013) From essential to persistent genes: a functional approach to constructing synthetic life. *Trends Genet.*, **29**, 273–279.
33 Danchin, A. and Sekowska, A. (2014) The logic of metabolism and its fuzzy consequences. *Environ. Microbiol.*, **16**, 19–28.
34 Wachtershauser, G. (2007) On the chemistry and evolution of the pioneer organism. *Chem. Biodivers.*, **4**, 584–602.

35 Sekowska, A., Kung, H.F., and Danchin, A. (2000) Sulfur metabolism in *Escherichia coli* and related bacteria: facts and fiction. *J. Mol. Microbiol. Biotechnol.*, **2**, 145–177.

36 Westheimer, F.H. (1987) Why nature chose phosphates. *Science*, **235**, 1173–1178.

37 Couture, R.M., Sekowska, A., Fang, G., and Danchin, A. (2012) Linking selenium biogeochemistry to the sulfur-dependent biological detoxification of arsenic. *Environ. Microbiol.*, **14**, 1612–1623.

38 Noria, S. and Danchin, A. (2002) Just so genome stories: what does my neighbor tell me? in *Uehara Memorial Foundation Symposium: Genome Science: Towards a New Paradigm?* (eds H. Yoshikawa, N. Ogasawara, and N. Satoh), Elsevier Science BV, Tokyo, pp. 3–13.

39 Nitschke, P., Guerdoux-Jamet, P., Chiapello, H., Faroux, G. *et al.* (1998) Indigo: a World-Wide-Web review of genomes and gene functions. *FEMS Microbiol. Rev.*, **22**, 207–227.

40 Rocha, E.P. and Danchin, A. (2002) Base composition bias might result from competition for metabolic resources. *Trends Genet.*, **18**, 291–294.

41 Danchin, A. (2009) A phylogenetic view of bacterial ribonucleases. *Prog. Mol. Biol. Transl. Sci.*, **85**, 1–41.

42 Chen, S., Wang, L., and Deng, Z. (2011) Twenty years hunting for sulfur in DNA. *Protein Cell*, **1**, 14–21.

43 Xie, X., Liang, J., Pu, T., Xu, F. *et al.* (2012) Phosphorothioate DNA as an antioxidant in bacteria. *Nucleic Acids Res.*, **40**, 9115–9124.

44 Schmidt, M. (2010) Xenobiology: a new form of life as the ultimate biosafety tool. *Bioessays*, **32**, 322–331.

45 Khudyakov, I.Y., Kirnos, M.D., Alexandrushkina, N.I., and Vanyushin, B.F. (1978) Cyanophage S-2L contains DNA with 2,6-diaminopurine substituted for adenine. *Virology*, **88**, 8–18.

46 Hutinet, G., Swarjo, M.A., and de Crecy-Lagard, V. (2016) Deazaguanine derivatives, examples of crosstalk between RNA and DNA modification pathways. *RNA Biol.*, 1–10.

47 Collier, J. (2009) Epigenetic regulation of the bacterial cell cycle. *Curr. Opin. Microbiol.*, **12**, 722–729.

48 Machnicka, M.A., Milanowska, K., Osman Oglou, O., Purta, E. *et al.* (2013) MODOMICS: a database of RNA modification pathways – 2013 update. *Nucleic Acids Res.*, **41**, D262–D267.

49 Akashi, H. and Gojobori, T. (2002) Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 3695–3700.

50 Chaplin, M. (2006) Do we underestimate the importance of water in cell biology? *Nat. Rev. Mol. Cell Biol.*, **7**, 861–866.

51 Danchin, A. and Fang, G. (2016) Unknown unknowns: essential genes in quest for function. *Microb. Biotechnol.*, **9**, 530–540.

52 Jakubowski, H. and Fersht, A.R. (1981) Alternative pathways for editing non-cognate amino acids by aminoacyl-tRNA synthetases. *Nucleic Acids Res.*, **9**, 3105–3117.

53 Budisa, N., Steipe, B., Demange, P., Eckerskorn, C. *et al.* (1995) High-level biosynthetic substitution of methionine in proteins by its analogs 2-aminohexanoic acid, selenomethionine, telluromethionine and ethionine in *Escherichia coli*. *Eur. J. Biochem.*, **230**, 788–796.

54 Rubino, J.T. and Franz, K.J. (2012) Coordination chemistry of copper proteins: how nature handles a toxic cargo for essential function. *J. Inorg. Biochem.*, **107**, 129–143.

55 Chan, C.M., Danchin, A., Marlière, P., and Sekowska, A. (2014) Paralogous metabolism: S-alkyl-cysteine degradation in *Bacillus subtilis*. *Environ. Microbiol.*, **16**, 101–117.

56 Koc, E.C. and Koc, H. (2012) Regulation of mammalian mitochondrial translation by post-translational modifications. *Biochim. Biophys. Acta*, **1819**, 1055–1066.

57 Schilling, B., Christensen, D., Davis, R., Sahu, A.K. *et al.* (2015) Protein acetylation dynamics in response to carbon overflow in *Escherichia coli*. *Mol. Microbiol.*, **98**, 847–863.

58 Martinez-Rodriguez, S., Martinez-Gomez, A.I., Rodriguez-Vico, F., Clemente-Jimenez, J.M. *et al.* (2010) Natural occurrence and industrial applications of d-amino acids: an overview. *Chem. Biodivers.*, **7**, 1531–1548.

59 Zhu, Y. and Lin, E.C. (1987) Loss of aldehyde dehydrogenase in an *Escherichia coli* mutant selected for growth on the rare sugar l-galactose. *J. Bacteriol.*, **169**, 785–789.

60 Danchin, A. and Sekowska, A. (2015) The logic of metabolism. *Perspect. Sci.*, **6**, 15–26.

61 Montanaro, L., Poggi, A., Visai, L., Ravaioli, S. *et al.* (2011) Extracellular DNA in biofilms. *Int. J. Artif. Organs*, **34**, 824–831.

62 Felczykowska, A., Bloch, S.K., Nejman-Falenczyk, B., and Baranska, S. (2012) Metagenomic approach in the investigation of new bioactive compounds in the marine environment. *Acta Biochim. Pol.*, **59**, 501–505.

63 Qin, S., Xing, K., Jiang, J.H., Xu, L.H. *et al.* (2011) Biodiversity, bioactive natural products and biotechnological potential of plant-associated endophytic actinobacteria. *Appl. Microbiol. Biotechnol.*, **89**, 457–473.

64 Zambare, V.P. and Christopher, L.P. (2011) Biopharmaceutical potential of lichens. *Pharm. Biol.*, **50**, 778–798.

65 Bittencourt, D., Oliveira, P.F., Prosdocimi, F., and Rech, E.L. (2012) Protein families, natural history and biotechnological aspects of spider silk. *Genet. Mol. Res.*, **11**, 2360–2380.

66 Colliec-Jouault, S., Bavington, C., and Delbarre-Ladrat, C. (2012) Heparin-like entities from marine organisms. *Handb. Exp. Pharmacol.*, 423–449.

67 Poulose, S., Panda, T., Nair, P.P., and Theodore, T. (2014) Biosynthesis of silver nanoparticles. *J. Nanosci. Nanotechnol.*, **14**, 2038–2049.

68 van Leeuwen, B.N., van der Wulp, A.M., Duijnstee, I., van Maris, A.J. *et al.* (2012) Fermentative production of isobutene. *Appl. Microbiol. Biotechnol.*, **93**, 1377–1387.

69 Danchin, A. (2017) Coping with inevitable accidents in metabolism. *Microb. Biotechnol.*, **10**, 57–72.

70 Achbergerova, L. and Nahalka, J. (2011) Polyphosphate – an ancient energy source and active metabolic regulator. *Microb. Cell Fact.*, **10**, 63.

71 Brock, J., Rhiel, E., Beutler, M., Salman, V. *et al.* (2012) Unusual polyphosphate inclusions observed in a marine *Beggiatoa* strain. *Antonie Van Leeuwenhoek*, **101**, 347–357.

72 Nikel, P.I., Chavarria, M., Martinez-Garcia, E., Taylor, A.C. *et al.* (2013) Accumulation of inorganic polyphosphate enables stress endurance and catalytic vigour in *Pseudomonas putida* KT2440. *Microb. Cell Fact.*, **12**, 50.

73 Friz, C.T. (1968) The biochemical composition of the free-living amoebae *Chaos chaos*, *Amoeba dubia* and *Amoeba proteus*. *Comp. Biochem. Physiol.*, **26**, 81–90.

74 Hallenbeck, P.C., Abo-Hashesh, M., and Ghosh, D. (2012) Strategies for improving biological hydrogen production. *Bioresour. Technol.*, **110**, 1–9.

75 Lakaniemi, A.M., Tuovinen, O.H., and Puhakka, J.A. (2013) Anaerobic conversion of microalgal biomass to sustainable energy carriers – a review. *Bioresour. Technol.*, **135**, 222–231.

76 van Niel, E.W. (2016) Biological processes for hydrogen production. *Adv. Biochem. Eng. Biotechnol.*, **156**, 155–193.

77 Logan, B.E. and Rabaey, K. (2012) Conversion of wastes into bioelectricity and chemicals by using microbial electrochemical technologies. *Science*, **337**, 686–690.

78 Wei, N., Quarterman, J., and Jin, Y.S. (2013) Marine macroalgae: an untapped resource for producing fuels and chemicals. *Trends Biotechnol.*, **31**, 70–77.

79 Saratale, G.D., Saratale, R.G., Shahid, M.K., Zhen, G. *et al.* (2017) A comprehensive overview on electro-active biofilms, role of exo-electrogens and their microbial niches in microbial fuel cells (MFCs). *Chemosphere*, **178**, 534–547.

80 Lovley, D.R. (2012) Long-range electron transport to Fe(III) oxide via pili with metallic-like conductivity. *Biochem. Soc. Trans.*, **40**, 1186–1190.

81 Malvankar, N.S. and Lovley, D.R. (2012) Microbial nanowires: a new paradigm for biological electron transfer and bioelectronics. *ChemSusChem*, **5**, 1039–1046.

82 Sure, S.K., Ackland, L.M., Torriero, A.A., Adholeya, A. *et al.* (2016) Microbial nanowires: an electrifying tale. *Microbiology*, **162**, 2017–2028.

83 Hashimoto, M., Ichimura, T., Mizoguchi, H., Tanaka, K. *et al.* (2005) Cell size and nucleoid organization of engineered *Escherichia coli* cells with a reduced genome. *Mol. Microbiol.*, **55**, 137–149.

84 Danchin, A., Guerdoux-Jamet, P., Moszer, I., and Nitschké, P. (2000) Mapping the bacterial cell architecture into the chromosome. *Philos. Trans. R. Soc. London, Ser. B*, **355**, 179–190.

85 Fang, G., Rocha, E.P., and Danchin, A. (2008) Persistence drives gene clustering in bacterial genomes. *BMC Genomics*, **9**, 4.

86 Jeong, K.S., Ahn, J., and Khodursky, A.B. (2004) Spatial patterns of transcriptional activity in the chromosome of *Escherichia coli*. *Genome Biol.*, **5**, R86.

87 Bailly-Bechet, M., Danchin, A., Iqbal, M., Marsili, M. *et al.* (2006) Codon usage domains over bacterial chromosomes. *PLoS Comput. Biol.*, **2**, e37.

88 Worcel, A. and Burgi, E. (1972) On the structure of the folded chromosome of *Escherichia coli*. *J. Mol. Biol.*, **71**, 127–147.

89 Dupaigne, P., Tonthat, N.K., Espeli, O., Whitfill, T. *et al.* (2012) Molecular basis for a protein-mediated DNA-bridging mechanism that functions in condensation of the *E. coli* chromosome. *Mol. Cell*, **48**, 560–571.

90 Rocha, E.P. (2008) The organization of the bacterial genome. *Annu. Rev. Genet.*, **42**, 211–233.

91 Cohanim, A.B., Trifonov, E.N., and Kashi, Y. (2006) Specific selection pressure at the third codon positions: contribution to 10- to 11-base periodicity in prokaryotic genomes. *J. Mol. Evol.*, **63**, 393–400.

92 Tomita, M., Wada, M., and Kawashima, Y. (1999) ApA dinucleotide periodicity in prokaryote, eukaryote, and organelle genomes. *J. Mol. Evol.*, **49**, 182–192.

93 Fukushima, A., Ikemura, T., Kinouchi, M., Oshima, T. *et al.* (2002) Periodicity in prokaryotic and eukaryotic genomes identified by power spectrum analysis. *Gene*, **300**, 203–211.

94 Herzel, H., Weiss, O., and Trifonov, E.N. (1999) 10–11 bp periodicities in complete genomes reflect protein structure and DNA folding. *Bioinformatics*, **15**, 187–193.

95 Larsabal, E. and Danchin, A. (2005) Genomes are covered with ubiquitous 11 bp periodic patterns, the "class A flexible patterns". *BMC Bioinf.*, **6**, 206.

96 Cho, B.K., Knight, E.M., Barrett, C.L., and Palsson, B.O. (2008) Genome-wide analysis of Fis binding in *Escherichia coli* indicates a causative role for A-/AT-tracts. *Genome Res.*, **18**, 900–910.

97 Gordon, B.R., Li, Y., Cote, A., Weirauch, M.T. *et al.* (2011) Structural basis for recognition of AT-rich DNA by unrelated xenogeneic silencing proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 10690–10695.

98 Nikolova, E.N., Bascom, G.D., Andricioaei, I., and Al-Hashimi, H.M. (2012) Probing sequence-specific DNA flexibility in a-tracts and pyrimidine-purine steps by nuclear magnetic resonance (13)C relaxation and molecular dynamics simulations. *Biochemistry*, **51**, 8654–8664.

99 Audit, B. and Ouzounis, C.A. (2003) From genes to genomes: universal scale-invariant properties of microbial chromosome organisation. *J. Mol. Biol.*, **332**, 617–633.

100 Willenbrock, H. and Ussery, D.W. (2004) Chromatin architecture and gene expression in *Escherichia coli*. *Genome Biol.*, **5**, 252.

101 Libby, E.A., Roggiani, M., and Goulian, M. (2012) Membrane protein expression triggers chromosomal locus repositioning in bacteria. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 7445–7450.

102 Singleton, C., Howard, T.P., and Smirnoff, N. (2014) Synthetic metabolons for metabolic engineering. *J. Exp. Bot.*, **65**, 1947–1954.

103 Murakawa, G.J., Kwan, C., Yamashita, J., and Nierlich, D.P. (1991) Transcription and decay of the lac messenger: role of an intergenic terminator. *J. Bacteriol.*, **173**, 28–36.

104 Saberi, S. and Emberly, E. (2010) Chromosome driven spatial patterning of proteins in bacteria. *PLoS Comput. Biol.*, **6**, e1000986.

105 Mingorance, J., Tamames, J., and Vicente, M. (2004) Genomic channeling in bacterial cell division. *J. Mol. Recognit.*, **17**, 481–487.

106 Tamames, J., Gonzalez-Moreno, M., Mingorance, J., Valencia, A. *et al.* (2001) Bringing gene order into bacterial shape. *Trends Genet.*, **17**, 124–126.

107 Macvanin, M. and Adhya, S. (2012) Architectural organization in *E. coli* nucleoid. *Biochim. Biophys. Acta*, **1819**, 830–835.
108 Rocha, E.P., Guerdoux-Jamet, P., Moszer, I., Viari, A. *et al.* (2000) Implication of gene distribution in the bacterial chromosome for the bacterial cell factory. *J. Biotechnol.*, **78**, 209–219.
109 Esnault, E., Valens, M., Espeli, O., and Boccard, F. (2007) Chromosome structuring limits genome plasticity in *Escherichia coli*. *PLos Genet.*, **3**, e226.
110 Touchon, M., Hoede, C., Tenaillon, O., Barbe, V. *et al.* (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLos Genet.*, **5**, e1000344.
111 Zhou, S., Kile, A., Bechner, M., Place, M. *et al.* (2004) Single-molecule approach to bacterial genomic comparisons via optical mapping. *J. Bacteriol.*, **186**, 7773–7782.
112 Rocha, E.P., Sekowska, A., and Danchin, A. (2000) Sulphur islands in the *Escherichia coli* genome: markers of the cell's architecture? *FEBS Lett.*, **476**, 8–11.
113 Espeli, O., Mercier, R., and Boccard, F. (2008) DNA dynamics vary according to macrodomain topography in the *E. coli* chromosome. *Mol. Microbiol.*, **68**, 1418–1427.
114 Thanbichler, M., Wang, S.C., and Shapiro, L. (2005) The bacterial nucleoid: a highly organized and dynamic structure. *J. Cell. Biochem.*, **96**, 506–521.
115 Badrinarayanan, A., Le, T.B., and Laub, M.T. (2015) Bacterial chromosome organization and segregation. *Annu. Rev. Cell Dev. Biol.*, **31**, 171–199.
116 Morrow, J.D. and Cooper, V.S. (2012) Evolutionary effects of translocations in bacterial genomes. *Genome Biol. Evol.*, **4**, 1256–1262.
117 Rocha, E.P. and Danchin, A. (2003) Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res.*, **31**, 6570–6577.
118 Carballido-Lopez, R. and Formstone, A. (2007) Shape determination in *Bacillus subtilis*. *Curr. Opin. Microbiol.*, **10**, 611–616.
119 Badrinarayanan, A., Lesterlin, C., Reyes-Lamothe, R., and Sherratt, D. (2012) The *Escherichia coli* SMC complex, MukBEF, shapes nucleoid organization independently of DNA replication. *J. Bacteriol.*, **194**, 4669–4676.
120 Gouffi, K., Gerard, F., Santini, C.L., and Wu, L.F. (2004) Dual topology of the *Escherichia coli* TatA protein. *J. Biol. Chem.*, **279**, 11608–11615.
121 Rapp, M., Granseth, E., Seppala, S., and von Heijne, G. (2006) Identification and evolution of dual-topology membrane proteins. *Nat. Struct. Mol. Biol.*, **13**, 112–116.
122 Danchin, A. (2009) Cells need safety valves. *Bioessays*, **31**, 769–773.
123 Pliotas, C. and Naismith, J.H. (2016) Spectator no more, the role of the membrane in regulating ion channel function. *Curr. Opin. Struct. Biol.*, **45**, 59–66.
124 Ellis, R.J. (2001) Macromolecular crowding: obvious but underappreciated. *Trends Biochem. Sci*, **26**, 597–604.
125 Spitzer, J. (2011) From water and ions to crowded biomacromolecules: in vivo structuring of a prokaryotic cell. *Microbiol. Mol. Biol. Rev.*, **75**, 491–506.
126 de Lorenzo, V., Sekowska, A., and Danchin, A. (2014) Chemical reactivity drives spatiotemporal organization of bacterial metabolism. *FEMS Microbiol. Rev.*, **39**, 96–119.

127 Dame, R.T., Tark-Dame, M., and Schiessel, H. (2011) A physical approach to segregation and folding of the *Caulobacter crescentus* genome. *Mol. Microbiol.*, **82**, 1311–1315.

128 Hornus, S., Levy, B., Lariviere, D., and Fourmentin, E. (2013) Easy DNA modeling and more with GraphiteLifeExplorer. *PLoS One*, **8**, e53609.

129 Takebayashi, S., Ryba, T., and Gilbert, D.M. (2012) Developmental control of replication timing defines a new breed of chromosomal domains with a novel mechanism of chromatin unfolding. *Nucleus*, **3**, 500–507.

130 Chen, H., Meisburger, S.P., Pabit, S.A., Sutton, J.L. *et al.* (2011) Ionic strength-dependent persistence lengths of single-stranded RNA and DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 799–804.

131 Hubner, M.R. and Spector, D.L. (2010) Chromatin dynamics. *Annu. Rev. Biophys.*, **39**, 471–489.

132 Jouffe, C., Cretenet, G., Symul, L., Martin, E. *et al.* (2013) The circadian clock coordinates ribosome biogenesis. *PLoS Biol.*, **11**, e1001455.

133 Nayak, M.K., Kulkarni, P.P., and Dash, D. (2013) Regulatory role of proteasome in determination of platelet life span. *J. Biol. Chem.*, **288**, 6826–6834.

134 Weintraub, S.J. and Deverman, B.E. (2007) Chronoregulation by asparagine deamidation. *Sci. STKE*, **2007**, re7.

135 Liu, C., Fu, X., Liu, L., Ren, X. *et al.* (2011) Sequential establishment of stripe patterns in an expanding cell population. *Science*, **334**, 238–241.

136 Binder, P.M. and Danchin, A. (2011) Life's demons: information and order in biology. What subcellular machines gather and process the information necessary to sustain life? *EMBO Rep.*, **12**, 495–499.

137 Budovsky, A., Fraifeld, V.E., and Aronov, S. (2010) Linking cell polarity, aging and rejuvenation. *Biogerontology*, **12**, 167–175.

138 Li, S., Ou, X.H., Wei, L., Wang, Z.B. *et al.* (2012) Septin 7 is required for orderly meiosis in mouse oocytes. *Cell Cycle*, **11**, 3211–3218.

139 Balakrishnan, L. and Bambara, R.A. (2013) Okazaki fragment metabolism. *Cold Spring Harb. Perspect. Biol.*, **5**, a010173.

140 Smith, G.R. (2012) How RecBCD enzyme and Chi promote DNA break repair and recombination: a molecular biologist's view. *Microbiol. Mol. Biol. Rev.*, **76**, 217–228.

141 Rodrigo-Brenni, M.C. and Hegde, R.S. (2012) Design principles of protein biosynthesis-coupled quality control. *Dev. Cell*, **23**, 896–907.

142 Prat, L., Heinemann, I.U., Aerni, H.R., Rinehart, J. *et al.* (2012) Carbon source-dependent expansion of the genetic code in bacteria. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 21070–21075.

143 Campbell, J.H., O'Donoghue, P., Campbell, A.G., Schwientek, P. *et al.* (2013) UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 5540–5545.

144 Bonnet, J., Yin, P., Ortiz, M.E., Subsoontorn, P. *et al.* (2013) Amplifying genetic logic gates. *Science*, **340**, 599–603.

145 Jung, K., Fried, L., Behr, S., and Heermann, R. (2012) Histidine kinases and response regulators in networks. *Curr. Opin. Microbiol.*, **15**, 118–124.

146 Follonier, S., Escapa, I.F., Fonseca, P.M., Henes, B. *et al.* (2013) New insights on the reorganization of gene transcription in *Pseudomonas putida* KT2440 at elevated pressure. *Microb. Cell Fact.*, **12**, 30.

147 Berthoumieux, S., de Jong, H., Baptist, G., Pinel, C. *et al.* (2013) Shared control of gene expression in bacteria by transcription factors and global physiology of the cell. *Mol. Syst. Biol.*, **9**, 634.

148 Wang, B. and Buck, M. (2012) Customizing cell signaling using engineered genetic logic circuits. *Trends Microbiol.*, **20**, 376–384.

149 Ang, J. and McMillen, D.R. (2013) Physical constraints on biological integral control design for homeostasis and sensory adaptation. *Biophys. J.*, **104**, 505–515.

Part II

Parts and Devices Supporting Control of Protein Expression and Activity

6

Constitutive and Regulated Promoters in Yeast: How to Design and Make Use of Promoters in *S. cerevisiae*

Diana S. M. Ottoz[1,2] *and Fabian Rudolf*[2]

[1] *ETH Zurich, Department of Biosystems Science and Engineering, Mattenstrasse 26, 4058 Basel, Switzerland*
[2] *Yale University, Department of Molecular Biophysics and Biochemistry, 333 Cedar street SHM C-111, New Haven, CT, 06520, USA*

The implementation of synthetic genetic circuits requires precise control of the expression of every gene involved. This can be achieved by choosing promoters that appropriately modulate transcription initiation in terms of intensity and duration in response to specific stimuli. In nature, promoters couple gene expression to the internal status of the cell and to the external conditions of the environment. Here, we describe *Saccharomyces cerevisiae* promoters. The characterization of the structural and functional features of natural promoters has been crucial for their application. Moreover, this knowledge led to the implementation of synthetic promoters displaying novel regulatory properties.

6.1 Introduction

The characterization of *Saccharomyces cerevisiae* promoters began by using them to drive expression of reporter genes. Systematic truncations and deletions of the promoter region of these constructs revealed that yeast promoters share a common modular structure [1, 2]. Each module has a defined role in the stimulation and regulation of transcription initiation [3, 4]. The characterization of natural promoters allowed their use in controlling the expression of heterologous genes [5–7].

Today, a large selection of well-characterized natural promoters is routinely exploited for controlling transcription in yeast [8, 9]. Although these promoters span a wide range of transcription initiation efficiencies, they usually do not cover them homogeneously; that is, most promoters display either very weak or very strong activity. Moreover, natural yeast promoters cannot be used to build orthogonal systems, since they are intimately linked to metabolism. These two limitations are overcome by constructing synthetic promoters, whose strength can be finely tuned and whose regulation can be independent of the metabolism.

Synthetic Biology: Parts, Devices and Applications, First Edition. Edited by Christina Smolke.

In this chapter, we deal with natural and synthetic promoters frequently used in yeast. After giving an overview of the essential features of natural promoters, we describe principles and strategies exploited to produce synthetic promoters and their cognate transcription factors. We leave out from the discussion other aspects of gene expression regulation, like gene copy number, transcription elongation and termination, transcript processing, mRNA stability, translation, and protein stability.

6.2 Yeast Promoters

A promoter is a DNA sequence enabling and regulating transcription initiation. In this section, we point out the essential structural and functional features of yeast promoters. For more detailed descriptions, reviews are available [10–14].

Yeast promoters consist of two functionally and physically distinguishable regions: the core promoter and the upstream element [4, 15, 16] (Figure 6.1, top). The core promoter is the region that carries the minimal information needed to start transcription, independently of any regulation [3, 15, 17]. It

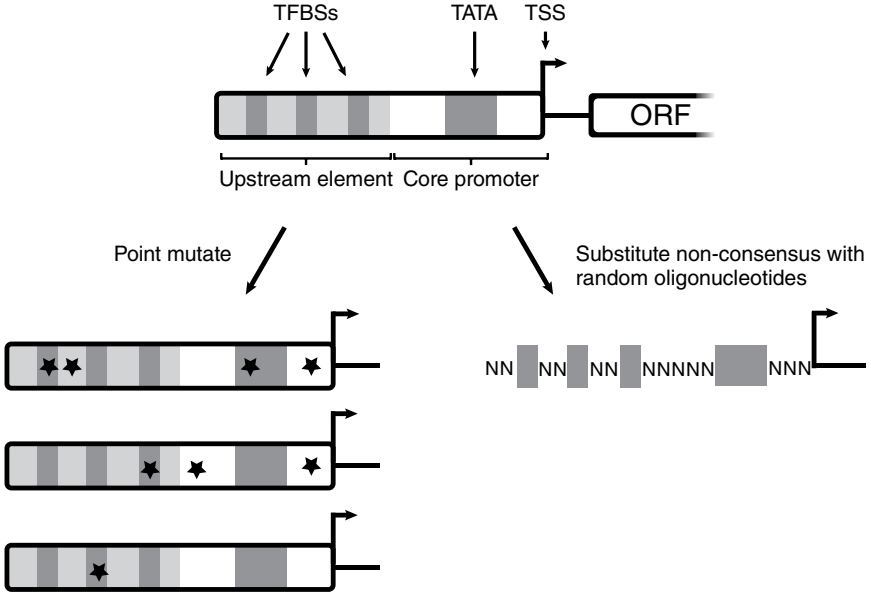


Figure 6.1 Modified natural yeast promoters. Top, typical bipartite structure of yeast promoters. Bottom left, promoter libraries obtained by point mutation. Random point mutations, illustrated as stars, are introduced by error-prone PCR along the sequence of the starting promoter. Bottom right, promoter libraries obtained by substituting non-consensus sequences with random oligonucleotides. By concentrating the mutations in the non-consensus regions, it is possible to fine-tune the strength of the starting promoter. N: nucleotide; ORF: open reading frame; TATA: TATA element; TFBS: transcription factor binding site; TSS: transcription initiation start site.

provides transcription initiation start sites (TSSs) defined by the consensus sequence $A(A_{rich})_5NPy\underline{A}(A/T)NN(A_{rich})_6$. In this consensus, $\underline{A}$ is the first transcribed nucleotide, N can be any nucleotide, and Py can be only a pyrimidine [18, 19]. As most yeast core promoters contain several TSSs, several transcript isoforms are usually produced from each promoter [20]. The core promoter is a platform for RNA polymerase II recruitment [21]. RNA polymerase II requires the assistance of general transcription factors to bind the promoter and become competent for transcription initiation. The general transcription factor called TATA-binding protein (TBP) recognizes the TATA element, a DNA sequence enriched for T and A [22], placed at variable distances upstream of the TSS(s) [18, 19]. The interaction between TBP and the TATA element triggers the stepwise recruitment of RNA polymerase II and other general transcription factors at the core promoter [22]. This results in the formation of the pre-initiation complex (PIC) , which is necessary to start transcription [12, 14]. Since the interaction between TBP and the TATA element directly triggers PIC assembly, the strength of this binding influences the overall transcription initiation efficiency. Therefore, the TATA element can be considered as a module acting as a scaling factor within the core promoter: strong TATA elements result in strong promoters; weak TATA elements result in weak promoters [23]. After PIC formation, RNA polymerase II searches the TSS(s) by scanning the template strand [24]. While RNA synthesis is not required for the scanning process, the selection of the TSS requires transcription. Indeed, limiting RNA polymerase II function leads to selection of TSS(s) further downstream [25]. The region between the TATA element and the TSS(s) is usually enriched in Ts, while downstream of the TSS(s), A is the preponderant nucleotide [26]. In strong promoters this biased nucleotide distribution is more evident than in weak promoters, suggesting that this feature could have an influence on transcription initiation efficiency by probably facilitating the identification of the TSS(s) during scanning [27].

The upstream element confers regulation by recruiting transcription factors [13]. Elements stimulating transcription initiation are called upstream activation sequences (UASs) and have some common features. First, their regulation depends on physiological stimuli [3, 16]. Second, their orientation does not affect their performance [28, 29]. Third, the distance between UASs and core promoter does not usually influence transcription initiation frequency [4, 28]. Fourth, UASs do not regulate transcription when they are placed downstream of the TATA element [28].

The observation that the UAS orientation and the distance from the TATA element do not influence transcription suggests that the UAS activity is independent of the core promoter; that is, these two regions do not interact with the same sets of proteins [30]. The distinct roles of core promoter and upstream element were demonstrated by constructing the first synthetic hybrid promoter, where the original UAS of a promoter was substituted with one of a second promoter. The resulting construct initiated transcription from the natural TSS of the first promoter but showed the typical regulation of the second [3]. This independence underscores the modular structure of yeast promoters, suggesting the possibility to combine several upstream elements. The resulting promoter reacts

to several physiological stimuli, converging transcription initiation on the TSS(s) defined by the core promoter [28, 29].

Transcription factors bind the upstream element on specific and well-defined DNA motifs called transcription factor binding sites (TFBSs). TFBSs are necessary and sufficient to confer regulation to a promoter [31].

In yeast, the most frequently observed mechanism of transcription initiation stimulation is activation by recruitment [32]. Transcription activators bind TFBSs located in the UAS. Their role merely consists of indicating the DNA region that needs to be transcribed. The binding of the transcription activator to its TFBS triggers the recruitment of the coactivators SAGA and TFIID, which in turn localize TBP to the core promoter. This array of protein–protein interactions results in the PIC assembly.

An important hint about the mechanism of activation by recruitment comes from the observation that DNA-binding and transcription activation activities of yeast transcription activators are functionally and physically separable. In fact, yeast activators display a modular structure containing, among others, a DNA-binding domain and an activation domain [33]. Truncations retaining either the DNA-binding or activation portion fail to initiate transcription. However, reassociation of these two portions restores function [34–36].

The modular structure of yeast transcription activators implies possible regulation of the mechanism of activation by recruitment. Masking the DNA-binding or activation activity by protein–protein interactions results in the failure of transcription initiation. The activity of the general repressor complex Cyc8–Tup1 consists in binding and covering the activation domains of target transcription activators. This interaction causes transcription initiation inhibition, even though the transcription activator is bound to its TFBS. Unmasking the activation domain by abolishing the interactions with Cyc8–Tup1 results in the recruitment of the transcriptional machinery [37].

In eukaryotes, DNA is not directly accessible, since it is wrapped around histones to form nucleosomes (reviewed in [38]). Nucleosomes provide a general inhibitory function that reduces basal transcription initiation of all genes (reviewed in [39]). As histones have a general affinity for DNA, nucleosomes form at random positions along DNA [40]. DNA-binding proteins that recognize specific binding sites compete with histones to interact with DNA (reviewed in [41]). However, the specific interaction of a DNA-binding protein to its binding site produces a physical barrier on the DNA that forces nucleosomes to phase around this point [40, 42]. In some promoters, nucleosome phasing may have an indirect role in transcription initiation stimulation by enhancing the accessibility of the TFBS of the transcription activator [42]. After the binding of the transcription activator, the nucleosomes must be displaced to assemble the PIC and start transcription. Therefore, the transcriptional machinery recruits factors involved in nucleosome remodeling [11]. The efficiency of nucleosome clearance is influenced by the propensity of DNA to be wrapped into nucleosomes [43]. The homopolymeric dA:dT sequences frequently observed in the UASs interact weakly with the histones and therefore cause the inefficient formation of nucleosomes in the region. This results in easier nucleosome clearance and stronger transcription [44]. Therefore, composition, length, and

position of these sequences along promoters have a direct effect on nucleosome occupancy and by consequence on transcription initiation efficiency [45].

Besides sequences stimulating transcription initiation, some yeast promoters also carry upstream elements that inhibit the process. These are called upstream repression sequences (URSs) and contain TFBSs that bind transcription repressors [12]. Some mechanisms of repression are also based on recruitment. Here, the binding of the repressor attracts corepressors to the promoter, which block transcription initiation by recruiting chromatin remodelers to make DNA less accessible for PIC assembly or by preventing the transcriptional machinery from starting [46].

6.3 Natural Yeast Promoters

We can distinguish two classes of natural promoters: regulated and constitutive. A wide selection of these promoters is used today to control gene expression. Although natural promoters are popular, their use is frequently limited to special genetic backgrounds and/or growth conditions. Nevertheless, the lessons learned from nature are essential to create synthetic systems more suitable for biotechnology or synthetic biology applications.

6.3.1 Regulated Promoters

The activity of a regulated promoter is, in terms of both timing and intensity, specifically dependent on a well-characterized stimulus, for example, chemical or physical agent. In many cases the stimulus operates a single specific TFBS. The promoters depending on galactose, inorganic phosphate, or copper described below are interesting examples.

The most used regulated promoters belong to the *GAL* genes, involved in galactose catabolism. The mechanism of their regulation is well characterized and involves several players (reviewed in [47]). *GAL4* is the main regulator of the *GAL* circuit and encodes a transcription activator. In the absence of galactose, the inhibitor Gal80 binds Gal4, preventing its activity. In the presence of galactose, Gal4 is released, as Gal80 is sequestered in the cytoplasm. This triggers the transcription of the Gal4 targets, which include *GAL1*, *GAL7*, and *GAL10*, encoding the enzymes of the Leloir pathway, and regulators of the circuit, such as *GAL80*, *GAL2*, and *GAL3*. The autocatalytic nature of the *GAL* circuit gives a switch-like response to galactose. The interruption of the positive feedback loop controlling the expression of the galactose permease *GAL2* results in a linear response of *GAL* genes to increasing amounts of galactose. This allows the induction of *GAL* promoters at intermediate levels [48]. Some yeast strains carry an extensive deletion in the *TRP1* locus resulting in the truncation of the adjacent *GAL3* promoter [49]. In this background, induction of the *GAL* genes is not fast and efficient, because the levels of Gal3 are low [50].

Galactose induces transcription of *GAL1* and *GAL10* by more than 1000-fold [51, 52]. *GAL1* and *GAL10* are in close proximity on the genome and diverge in their orientation [52, 53]. Deletion analysis of the DNA sequence lying between

the two open reading frames revealed the presence of a galactose-dependent UAS [54, 55]. The UAS contains two shared TFBSs bound by the transcriptional activator Gal4. These TFBSs are sufficient to confer galactose-dependent regulation to a promoter [31]. The first yeast synthetic hybrid promoter, discussed above, was assembled by placing the *GAL1–10* UAS upstream of the core promoter region of another gene. The construct exhibited the typical *GAL1–10* galactose-dependent regulation [3].

The promoter of the acid phosphatase (*PHO5*) contains two homologous TFBSs bound by the transcription activators Pho4 and Pho2 when inorganic phosphate is depleted in the culture medium. The presence of inorganic phosphate in the medium leads to Pho4 sequestration in a protein complex that does not allow its binding to the *PHO5* promoter [56, 57]. Although this promoter displays some basal activity in the repressed state [56], it has been successfully used to express heterologous genes, like the hepatitis B surface antigen [58].

The promoters of genes involved in copper metabolism are frequently used for driving transcription of heterologous genes [59, 60]. The *CUP1* promoter is stimulated by copper [61]. Its activation depends on the transcription activator Ace1, whose ability to bind its TFBSs is controlled by its interaction with copper ions [62]. Although this is a popular promoter, its use is limited to strains carrying the wild-type *CUP1* locus. With this genetic background it is possible to avoid toxic effects related to excess copper, since the *CUP1* gene encodes a metallothionein acting as a copper chelator. However, copper is essential in biological processes like respiration; therefore it is usually present in traces in culture media. This small amount of copper causes a substantial basal expression of genes under the control of the *CUP1* promoter. The metallothionein encoded by the wild-type *CUP1* locus contributes to lowering the amount of available copper. As a consequence the basal activity of the heterologous construct is lowered [63]. An excess of copper prevents the transcription of genes encoding copper transporters, like *CTR1* and *CTR3* [64]. The promoters of these genes contain specific TFBSs bound by the transcription activator Mac1. When Mac1 interacts with copper ions, its DNA-binding and activation activities are inhibited [65]. A collection of expression vectors containing *CUP1*, *CTR1*, and *CTR3* promoters is available for coordinated induction and inhibition experiments [66].

Regulation of the promoters described so far depends on a single TFBS. However, some promoters display a combination of TFBSs bound by different transcription factors. This results in more sophisticated regulation. Several examples are described below.

The promoter of *DAN1*, a mannoprotein, contains a set of TFBSs bound by both activators and repressors. The combination of the activity of these transcription factors results in complete repression in the presence of oxygen and full activation when this gas is absent from the culture medium [67]. For induction, this promoter requires stringent anaerobiosis, which can be realized by bubbling nitrogen in the cultures. However, this experimental setup is not convenient for large scale overexpression experiments. As such, random mutagenesis of the *DAN1* promoter yielded variants less sensitive to oxygen that can be induced in microaerobiosis [68].

Carbon catabolite repression is the set of regulatory mechanisms forcing cells to preferentially use glucose and fructose over other carbon sources (reviewed in [69, 70]). For example, the *GAL* circuit is fully repressed by glucose, even when galactose is present in the culture medium [51]. Carbon catabolite repression affects the levels and activity of enzymes involved in energy metabolism. We can distinguish two main ways of transcription initiation repression by glucose. The first way is direct, when the presence of glucose triggers the recruitment of transcription repressors to the target promoters. Mig1 represses transcription initiation via Cyc8–Tup1. When glucose is depleted, Mig1 is phosphorylated by Snf1 and is consequently relocalized to the cytoplasm, thus abolishing repression of the target promoters [69]. A well-characterized target of Mig1 is *GAL1*, which contains the cognate TFBS upstream of its TATA element [54, 55]. The second way is indirect, when transcription repression is achieved by inactivating transcription activators. For example, Adr1 is inactive in the presence of glucose; therefore it cannot trigger transcription initiation [71]. A well-characterized natural target of this transcription activator is alcohol dehydrogenase 2 (*ADH2*), which is repressed when yeast is grown in glucose [72]. When the TFBS recognized by Adr1 is cloned in front of the fermentative alcohol dehydrogenase 1 (*ADH1*) gene, its expression shows catabolite repression [15]. Carbon catabolite repression can be exploited to repress genes of interest until glucose is depleted in the culture medium. Besides the *ADH2* promoter [73], that of *JEN1*, the main lactate and pyruvate transporter, is also used. The *JEN1* promoter was initially selected to construct biosensors for measuring sugar concentrations, since it reacts specifically to carbon sources and is insensitive to most types of cell stresses [74].

6.3.2 Constitutive Promoters

A constitutive promoter displays a relatively constant activity that is not significantly altered by stimuli. In most cases, the activity of constitutive promoters is coupled to the growth rate, which depends on the level of glucose, the preferred carbon source of yeast [69, 75, 76]. This constant transcription is ensured by a complex combination of TFBSs.

The most used constitutive promoters belong to genes involved in primary cell metabolism such as glycolysis and fermentation. The main reason for this selection is historical; mutations affecting these genes were relatively easily isolated and characterized. The regulation of the expression of glycolytic and fermentative enzymes takes place mainly at the transcriptional level and correlates with glucose concentration and growth curve stage [76–80]. The promoters of these genes share common TFBSs recognized by the regulators Rap1 and Gcr1 [29, 81, 82], which ensure coordinated transcription [83, 84]. The promoters of phosphoglycerate kinase 1 (*PGK1*) and glyceraldehyde-3-phosphate dehydrogenase 3 (*TDH3*) are among the strongest ones known [77, 80]. The promoter of the fermentative *ADH1* was one of the first used to overexpress a heterologous protein in yeast [5]. A vector containing this promoter was also used to produce the human hepatitis B vaccine [7]. Although considered strong, the *ADH1* promoter is weaker than those of *PGK1* or *TDH3* [9, 80].

The cytochrome c isoform 1 (*CYC1*) promoter has been extensively characterized [1, 16]. This gene is involved in cell respiration and its transcription is triggered by heme [16, 30]. When this metabolite is present in the cells, it binds the transcription activator Hap1, which can then recognize its cognate TFBS [30, 85]. Today, a truncated version of this promoter is used to drive mild and relatively constant expression in fermentative growth conditions. For this reason the *CYC1* promoter is usually described as a constitutive promoter [9].

Besides genes involved in energy metabolism, those involved in other basic tasks, like cell shape maintenance and translation, also have constitutive promoters. For example, the promoter of the gene encoding β -actin (*ACT1*) displays a combination of regulatory elements ensuring a constant transcription in both fermentative and non-fermentative growth conditions [86]. Similarly, the translation elongation factor EF-1 α (*TEF1*) has a promoter ensuring approximately stable expression during all growth phases and in media containing different carbon sources [76, 80, 87].

6.4 Synthetic Yeast Promoters

A synthetic promoter carries nonnative sequences. We describe two main groups of synthetic promoters. One includes modified versions of natural promoters, and the other contains hybrid promoters. As illustrated below, the main difference between promoters belonging to each class is the strategy used to construct them.

6.4.1 Modified Natural Promoters

The systematic modification of a promoter leads to a library spanning a wide range of transcription initiation frequencies. Within this library, each member drives transcription initiation with a specific strength. Since the members of the library are derived from a single promoter, they share similar regulatory features; that is, they respond to the same stimulus [88–90]. There are two main methods for obtaining promoter libraries: either by introducing point mutations or by substituting short sequences with randomized oligonucleotides (reviewed in [91]) (Figure 6.1).

Mutations in essential regulatory sequences are likely to cause a substantial change in activity, because they can alter the binding affinity of the cognate proteins. A library of *TEF1* promoter variants was obtained by error-prone PCR. With this approach the point mutations spanned along the complete sequence of the promoter [92]. The library covered a range of activities from 8% to 120% relative to the native *TEF1* promoter [93]. A variation of this strategy consists in limiting the point mutations to specific regions of the promoter, for example, to the TATA element. These modifications alter the efficiency of the PIC assembly [94]; therefore they affect the overall performance of the promoter [23].

An alternative strategy for obtaining promoter libraries is the substitution of the non-consensus sequences of the promoter with random sequences. Non-consensus sequences usually do not play a direct role in transcription initiation regulation; that is, they do not bind to specific proteins. However, those sequences might modulate the process indirectly, for example, by keeping the optimal distance between functional sequences [91], by influencing the local DNA helical parameters [95], or by modulating the efficiency of nucleosome formation and clearance [43, 96]. Therefore, promoter variants containing modifications in non-consensus sequences will differ from each other by small changes in strength. The modification strategy is based on the synthesis of libraries of oligonucleotides encoding the promoter sequence. In each oligonucleotide the consensus sequences are separated by degenerate stretches of nucleotides of variable length [88]. This approach was used to modify the profilin (*PFY1*) promoter. The library obtained spanned a range of activities from 11% to 100% relative to the starting promoter [89].

6.4.2 Synthetic Hybrid Promoters

Synthetic hybrid promoters combine DNA sequences originally belonging to different promoters, but retain the typical bipartite structure of natural promoters [91] (Figure 6.2).

The choice of the core promoter has effects on the overall performance of the hybrid promoter, as it controls the efficiency of the PIC assembly [23] and the identification of the TSS(s) [26]. Frequently, synthetic hybrid promoters contain the native core promoter of inducible genes, for example, *LEU2* or *CYC1* [3, 90, 97]. Strong synthetic core promoters have been isolated from DNA libraries where the TATA element and the TSS consensus sequence were separated by a randomized spacer of 30 nucleotides. An additional stretch of 30 nucleotides placed between the TATA element and the upstream TFBSs improves the core promoter robustness by possibly avoiding steric hindrances between the transcription factors, TBP, and other general transcription factors that bind to the TFBSs or the core promoter [98].

The main advantage of the synthetic hybrid promoter approach is the possibility of using any DNA sequence targeted by a protein as an upstream element. Endogenous TFBSs link the synthetic promoter to a regulatory pathway. For example, by placing the UAS of *GAL1–10* in front of the *TDH3* promoter, a hybrid promoter that is active in glucose and further stimulated by galactose was obtained [90]. Heterologous TFBSs either belong originally to other species or are artificial sequences. They enable the implementation of orthogonal transcription systems that are independent of metabolism [99]. In fact, heterologous sequences are not recognized by any yeast transcription factor, unless they are homologous or, by chance, similar to endogenous sequences [34, 97, 100]. Promoters containing combinations of TFBSs are sensitive to several stimuli [90].

Modification of the binding affinity between the transcription factor and its cognate TFBS results in promoter strength modulation [99, 101]. Alternatively, the strength of the promoter can be tuned by increasing the number of TFBS

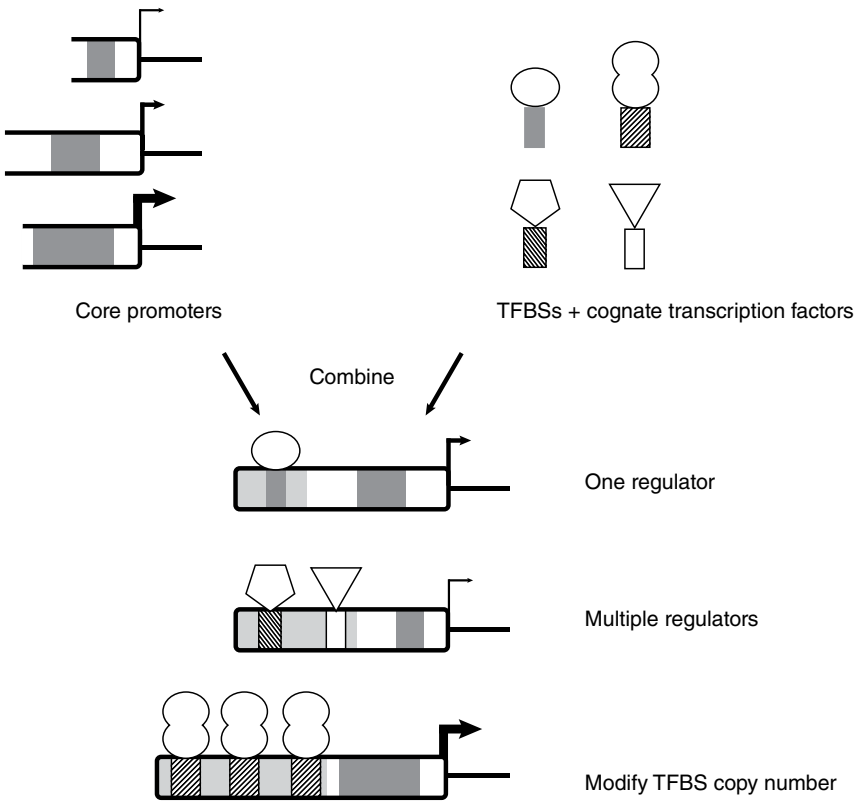


Figure 6.2 Synthetic hybrid promoters are obtained by combining a core promoter and one or more transcription factor binding sites (TFBSs). Each TFBS is specifically recognized by a transcription factor. By selecting the TFBSs, it is possible to choose which regulation the synthetic hybrid promoter should display. The combination of two or more different TFBSs results in a combinatorial regulation. The multiplication of the TFBS copy number results in the adjustment of the promoter strength.

copies; in most cases, a linear relationship is observed [90, 99, 102–104]. The activity of promoters also depends on the spacer sequences placed between regulatory elements [59, 88, 105, 106]. These sequences can have an impact on the efficiency of nucleosome clearance [43, 96]. Homopolymeric dA:dT or dG:dC stretches disfavor nucleosome formation, thereby increasing transcription initiation efficiency [44]. On the contrary, DNA sequences containing dA:dT dinucleotides alternating with dG:dC dinucleotides are wrapped very efficiently into nucleosomes, thereby inhibiting transcription initiation efficiency [43]. Therefore, it is possible to fine-tune initiation efficiency by modulating the length, composition, and location of dA:dT or dG:dC stretches within the promoter [44, 45]. Finally, it is also necessary to avoid the formation of structures that may have an unpredictable influence on the promoter performance. For example, placing a transcriptional terminator-like sequence between the upstream element and the core promoter can depress transcription initiation [107].

Hybrid promoters containing heterologous TFBSs need a heterologous transcription factor. To ensure orthogonality, this transcription factor should not have any other target than the heterologous promoter itself. The modular structure of natural transcription factors suggests that it is possible to combine different protein domains to obtain new factors.

A protein able to bind DNA can stimulate transcription when fused to an activation domain. The first heterologous transcription factor tested in yeast contained the bacterial DNA-binding protein LexA fused to an activation domain and triggered transcription of promoters containing LexA TFBSs [34, 108, 109]. Transcription activators containing the bacterial DNA-binding protein *tet*R are regulated by tetracycline [97, 110], which prevents the binding of *tet*R to the cognate TFBSs [111–113]; therefore, the expression of the target gene can be modulated by adjusting the concentration of this chemical in the culture medium. A reverse *tet*R mutant, which binds its TFBS upon addition of tetracycline, is also available. However, transcription activators containing reverse *tet*R have a relatively strong basal activity in the absence of tetracycline [114].

Hybrid promoters containing artificial TFBSs require the construction of artificial DNA-binding domains. Zinc fingers and transcriptional activator-like effectors (TALEs) are short peptidic modules binding to specific and short DNA sequences. Protein engineering has diversified these modules, and libraries of protein moieties recognizing virtually all DNA sequences of three to four nucleotides have been constructed. By fusing several zinc fingers or TALE modules, it is possible to obtain arrays that specifically bind longer DNA sequences [89, 115, 116]. Artificial transcription activators are obtained by fusing these DNA-binding domains to activation domains [100, 106].

In general, any mechanism able to target an activation domain to the DNA can be used to stimulate transcription initiation. In the clustered regularly interspaced palindromic repeats (CRISPR)-derived system, the target DNA sequences are identified via RNA-mediated interactions, instead of binding of a protein. In this system, the DNA-binding activity consists of a single-guide RNA (sgRNA), which targets specifically the DNA region to be regulated, and the catalytically inactive version of the protein Cas9 (dCas9), which binds specifically the sgRNA. By fusing dCas9 to an activation domain, a transcription activator is obtained [117, 118].

The activation domain stimulates transcription initiation by establishing protein–protein interactions with coactivators and components of the transcriptional machinery [119, 120]. While DNA-binding domains have well-defined conserved architectures (reviewed in [12]), activation domains do not share common structures, except for a marked acidity [35, 121]. They usually consist of multiple unstructured acidic patches; each acidic patch triggers transcription initiation when fused to a DNA-binding domain [122]. Any peptide stretch displaying such properties can be used to activate transcription [123].

The DNA-binding and activation activities do not need to reside on a unique protein but can be physically separated on two different molecules. An interaction between these two is sufficient for transcription initiation. This is exploited in the yeast two-hybrid assay [124, 125]. This principle was also used to construct a light switchable system. Here, the DNA-binding domain and the activation

domain of Gal4 are separated and fused to the chromoprotein phytochrome PhyB and its interactor Pif3, respectively. Red light converts the PhyB fusion into its active form, which interacts with the Pif3 fusion, activating transcription. Far-red light converts the PhyB fusion into its inactive form, which cannot bind Pif3, disabling transcription initiation [126].

The activity of heterologous transcription factors can be precisely controlled by fusing additional domains that, for example, trigger nuclear localization upon a specific stimulus. The human estrogen receptor, when fused to a transcription activator, confers a hormone-dependent regulation. This chimera triggers transcription initiation only when β -estradiol is added to the culture medium [102]. Binding of the hormone to the estrogen receptor causes the nuclear localization of the transcription activator, which, in the absence of inducer, is diffusing all over the cell [127]. An activator containing the Gal4 DNA-binding domain and the estrogen receptor binds *GAL* promoters, but its activity does not depend on the carbon source [127, 128]. Estrogen-regulated activators based on heterologous DNA-binding domains such as LexA or synthetic zinc fingers result in orthogonal systems that specifically regulate the expression of the target promoters [100, 102]. The LexA-based activator induces the expression of the target gene in different growth conditions. Its overall activity can be finely tuned with the concentration of β -estradiol in the culture medium, the number of LexA TFBSs in the target promoter, and the choice of the activation domain [102].

An essential aspect of regulated synthetic promoters is the tightness of their regulation. A promoter is tightly regulated when it does not have any basal activity in the absence of the stimulus. The basal activity of some promoters depends on the residual activity of the transcription activator in the absence of the stimulus [114]. Alternatively, the basal expression can be the consequence of ectopic transcriptional events starting upstream of the promoter itself [20]. In this case, the insulation of the synthetic transcription unit is necessary. This can be obtained by placing a transcriptional terminator in front of the synthetic promoter [129].

Regulation of gene expression can also be achieved by repression of transcription initiation. A protein binding the DNA between the upstream element and the core promoter or within the core promoter prevents the establishment of the interactions needed for the effective recruitment of the PIC, causing transcription repression by steric hindrance [33, 36, 89, 107, 130]. The DNA-binding protein *tet*R was used to systematically study this kind of repression. A collection of *GAL1* promoter variants containing different number of *tet*R TFBSs placed between the TATA element and the TSS was tested. It was observed that increasing the number of such TFBSs reduced the basal expression of the system. Moreover, repression was stronger when the TFBSs were placed in close proximity to the TATA element [103, 105]. At intermediate levels of induction, the expression levels of the genes targeted by *tet*R showed a broad cell-to-cell variability. Reduction of such a cell-to-cell variability was obtained by placing *tet*R expression under its own control, implementing a negative feedback loop [131]. *Tet*R expression under negative feedback control also resulted in a "linearized" dose–response curve, allowing for larger concentration ranges of tetracycline and therefore better titratability. This negative feedback-based concept has also

been applied to mammalian synthetic gene circuits [132]. Repression by steric hindrance can also be obtained by using a CRISPR-derived system. A complex consisting of sgRNA and dCas9 competes with the transcription activator of the target promoter, as it targets the same TFBS. The sgRNA–dCas9 complex prevents transcription initiation by blocking the access of the TFBS [117]. However, neither the *tet*R nor other bacterial repressor domains nor the CRISPR-based systems show a sufficiently tight repression of the basal level to be useful in a broad setting.

The basal level of heterologous repression systems can be further reduced by fusing a DNA-binding domain to a component of the eukaryotic transcriptional repressor complex, like Tup1 or Cyc8. LexA, when fused either to Tup1 or to Cyc8, mediates repression of hybrid promoters containing LexA TFBSs [133, 134]. The CRISPR-derived system has been used in a similar strategy. dCas9 was fused to a mammalian repressor that recruits a yeast histone deacetylase. This repressor was targeted to the *TEF1* promoter by designing a specific sgRNA [117]. As a drawback, these systems slow down the transcriptional induction kinetics and affect the expression levels of (endogenous) genes located in close proximity.

6.5 Conclusions

In this chapter, we highlighted some examples of both regulated and constitutive natural yeast promoters. The characterization of these sequences allowed for the identification of structural and functional features that are exploited to build synthetic promoters and heterologous transcription factors. Examples of the application of these promoters in synthetic biology have been reviewed in [75, 91, 135–137].

Today, in the context of implementation of novel functions in cells, the construction of robust promoters is crucial [99]. Recent efforts to transform biotechnology and synthetic biology into more engineering-like disciplines also motivate the construction of synthetic promoters. In fact, their implementation is an essential step for the abstraction and standardization of concepts like promoter structure and transcription initiation [138]. In this perspective, modularization and orthogonality are the aspects that need to be further developed.

Modularization enables the implementation of new promoters and transcription factors by combining well-characterized and structurally independent modules. Orthogonal systems do not depend on and influence the endogenous metabolism. This ensures a robust behavior and the possibility to reuse the system in different environments and contexts. Today, zinc finger, TALE, and CRISPR-based technologies allow the design of artificial transcription factors that recognize unique sequences [99, 100, 139]. Additionally, the CRISPR-based toolkits available now allow for simple construction of strains containing constructs with the different mechanisms discussed in this review [140]. Together, modularization and orthogonality ensure versatility and the possibility to easily construct new systems with improved or new functionalities.

Definitions

Transcription initiation: Initial steps of transcription until the formation of the first RNA bond.
Promoter: A DNA sequence enabling and regulating transcription initiation.
Transcription initiation start site (TSS): The first nucleotide of a DNA sequence to be transcribed into RNA.
Core promoter: Promoter region defining the TSS(s) and the assembly of the PIC.
Pre-initiation complex (PIC): Protein complex containing RNA polymerase II and the general transcription factors that assemble on the core promoter.
Transcription factor: A protein regulating transcription initiation. A transcription factor is not a subunit of RNA polymerase II.
Upstream element: Promoter region conferring regulation to transcription initiation. It contains transcription factor binding sites (TFBSs).
Orthogonal system: A system that is independent of cell physiology.

References

1 Guarente, L. and Ptashne, M. (1981) Fusion of *Escherichia coli* lacZ to the cytochrome c gene of *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U.S.A.*, **78** (4), 2199–2203.
2 Rose, M., Casadaban, M.J., and Botstein, D. (1981) Yeast genes fused to beta-galactosidase in *Escherichia coli* can be expressed normally in yeast. *Proc. Natl. Acad. Sci. U.S.A.*, **78** (4), 2460–2464.
3 Guarente, L., Yocum, R.R., and Gifford, P. (1982) A *GAL10-CYC1* hybrid yeast promoter identifies the *GAL4* regulatory region as an upstream site. *Proc. Natl. Acad. Sci. U.S.A.*, **79** (23), 7410–7414.
4 Struhl, K. (1982) The yeast *his3* promoter contains at least two distinct elements. *Proc. Natl. Acad. Sci. U.S.A.*, **79** (23), 7385–7389.
5 Hitzeman, R.A., Hagie, F.E., Levine, H.L., Goeddel, D.V., Ammerer, G., and Hall, B.D. (1981) Expression of a human gene for interferon in yeast. *Nature*, **293** (5835), 717–722.
6 Ramer, S.W., Elledge, S.J., and Davis, R.W. (1992) Dominant genetics using a yeast genomic library under the control of a strong inducible promoter. *Proc. Natl. Acad. Sci. U.S.A.*, **89** (23), 11 589–11 593.
7 McAleer, W.J., Buynak, E.B., Maigetter, R.Z., Wampler, D.E., Miller, W.J., and Hilleman, M.R. (1984) Human hepatitis B vaccine from recombinant yeast. *Nature*, **307** (5947), 178–180.
8 Mumberg, D., Müller, R., and Funk, M. (1994) Regulatable promoters of *Saccharomyces cerevisiae*: comparison of transcriptional activity and their use for heterologous expression. *Nucleic Acids Res.*, **22** (25), 5767–5768.
9 Mumberg, D., Müller, R., and Funk, M. (1995) Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds. *Gene*, **156** (1), 119–122.

10 Struhl, K. (1995) Yeast transcriptional regulatory mechanisms. *Annu. Rev. Genet.*, **29** (1), 651–674.
11 Weake, V.M. and Workman, J.L. (2010) Inducible gene expression: diverse regulatory mechanisms. *Nat. Rev. Genet.*, **11** (6), 426–437.
12 Hahn, S., Young, E.T., and Hinnebusch, A. (2011) Transcriptional regulation in *Saccharomyces cerevisiae*: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics*, **189** (3), 705–736.
13 Guarente, L. (1984) Yeast promoters: positive and negative elements. *Cell*, **36** (4), 799–800.
14 Sainsbury, S., Bernecky, C., and Cramer, P. (2015) Structural basis of transcription initiation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.*, **16** (3), 129–143.
15 Beier, D.R. and Young, E.T. (1982) Characterization of a regulatory region upstream of the *ADR2* locus of *S. cerevisiae*. *Nature*, **300** (5894), 724–728.
16 Guarente, L. and Mason, T. (1983) Heme regulates transcription of the *CYC1* gene of *S. cerevisiae* via an upstream activation site. *Cell*, **32** (4), 1279–1286.
17 Faye, G., Leung, D.W., Tatchell, K., Hall, B.D., and Smith, M. (1981) Deletion mapping of sequences essential for in vivo transcription of the iso-1-cytochrome c gene. *Proc. Natl. Acad. Sci. U.S.A.*, **78** (4), 2258–2262.
18 Hahn, S., Hoar, E.T., and Guarente, L. (1985) Each of three "TATA elements" specifies a subset of the transcription initiation sites at the *CYC-1* promoter of *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U.S.A.*, **82** (24), 8562–8566.
19 Zhang, Z. and Dietrich, F.S. (2005) Mapping of transcription start sites in *Saccharomyces cerevisiae* using 5' SAGE. *Nucleic Acids Res.*, **33** (9), 2838–2851.
20 Pelechano, V., Wei, W., and Steinmetz, L.M. (2013) Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature*, **497** (7447), 127–131.
21 Venters, B.J., Wachi, S., Mavrich, T.N., Andersen, B.E., Jena, P., Sinnamon, A.J., Jain, P., Rolleri, N.S., Jiang, C., Hemeryck-Walsh, C., and Pugh, B.F. (2011) A comprehensive genomic binding map of gene and chromatin regulatory proteins in *Saccharomyces*. *Mol. Cell*, **41** (4), 480–492.
22 Rhee, H.S. and Pugh, B.F. (2012) Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature*, **483** (7389), 295–301.
23 Mogno, I., Vallania, F., Mitra, R.D., and Cohen, B.A. (2010) TATA is a modular component of synthetic promoters. *Genome Res.*, **20** (10), 1391–1397.
24 Giardina, C. and Lis, J.T. (1993) DNA melting on yeast RNA polymerase II promoters. *Science*, **261** (5122), 759–762.
25 Fishburn, J., Galburt, E., and Hahn, S. (2016) Transcription start site scanning and the requirement for ATP during transcription initiation by RNA Polymerase II. *J. Biol. Chem.*, **291** (25), 13 040–13 047.
26 Maicas, E. and Friesen, J.D. (1990) A sequence pattern that occurs at the transcription initiation region of yeast RNA polymerase II promoters. *Nucleic Acids Res.*, **18** (11), 3387–3393.
27 Lubliner, S., Keren, L., and Segal, E. (2013) Sequence features of yeast and human core promoters that are predictive of maximal promoter activity. *Nucleic Acids Res.*, **41** (11), 5569–5581.

28 Guarente, L. and Hoar, E. (1984) Upstream activation sites of the *CYC1* gene of *Saccharomyces cerevisiae* are active when inverted but not when placed downstream of the "TATA box". *Proc. Natl. Acad. Sci. U.S.A.*, **81** (24), 7860–7864.

29 Ogden, J.E., Stanway, C., Kim, S., Mellor, J., Kingsman, A.J., and Kingsman, S.M. (1986) Efficient expression of the *Saccharomyces cerevisiae PGK* gene depends on an upstream activation sequence but does not require TATA sequences. *Mol. Cell. Biol.*, **6** (12), 4335–4343.

30 Guarente, L., Lalonde, B., Gifford, P., and Alani, E. (1984) Distinctly regulated tandem upstream activation sites mediate catabolite repression of the *CYC1* gene of *S. cerevisiae*. *Cell*, **36** (2), 503–511.

31 Giniger, E., Varnum, S.M., and Ptashne, M. (1985) Specific DNA binding of GAL4, a positive regulatory protein of yeast. *Cell*, **40** (4), 767–774.

32 Ptashne, M. and Gann, A. (1997) Transcriptional activation by recruitment. *Nature*, **386** (6625), 569–577.

33 Keegan, L., Gill, G., and Ptashne, M. (1986) Separation of DNA binding from the transcription-activating function of a eukaryotic regulatory protein. *Science*, **231** (4739), 699–704.

34 Brent, R. and Ptashne, M. (1985) A eukaryotic transcriptional activator bearing the DNA specificity of a prokaryotic repressor. *Cell*, **43** (3 Pt 2), 729–736.

35 Ma, J. and Ptashne, M. (1987) Deletion analysis of GAL4 defines two transcriptional activating segments. *Cell*, **48** (5), 847–853.

36 Hope, I.A. and Struhl, K. (1986) Functional dissection of a eukaryotic transcriptional activator protein, GCN4 of yeast. *Cell*, **46** (6), 885–894.

37 Wong, K.H. and Struhl, K. (2011) The Cyc8-Tup1 complex inhibits transcription primarily by masking the activation domain of the recruiting protein. *Genes Dev.*, **25** (23), 2525–2539.

38 Campos, E.I. and Reinberg, D. (2009) Histones: annotating chromatin. *Annu. Rev. Genet.*, **43** (1), 559–599.

39 Struhl, K. (1999) Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell*, **98** (1), 1–4.

40 Kornberg, R. (1981) The location of nucleosomes in chromatin: specific or statistical? *Nature*, **292** (5824), 579–580.

41 Wang, X., Bai, L., Bryant, G.O., and Ptashne, M. (2011) Nucleosomes and the accessibility problem. *Trends Genet.*, **27** (12), 487–492.

42 Floer, M., Wang, X., Prabhu, V., Berrozpe, G., Narayan, S., Spagna, D., Alvarez, D., Kendall, J., Krasnitz, A., Stepansky, A., Hicks, J., Bryant, G.O., and Ptashne, M. (2010) A RSC/nucleosome complex determines chromatin architecture and facilitates activator binding. *Cell*, **141** (3), 407–418.

43 Wang, X., Bryant, G.O., Floer, M., Spagna, D., and Ptashne, M. (2011) An effect of DNA sequence on nucleosome occupancy and removal. *Nat. Struct. Mol. Biol.*, **18** (4), 507–509.

44 Iyer, V. and Struhl, K. (1995) Poly(dA: dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J.*, **14** (11), 2570–2579.

45 Raveh-Sadka, T., Levo, M., Shabi, U., Shany, B., Keren, L., Lotan-Pompan, M., Zeevi, D., Sharon, E., Weinberger, A., and Segal, E. (2012) Manipulating

nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat. Genet.*, **44** (7), 743–750.

46 Kassir, Y., Adir, N., Boger-Nadjar, E., Raviv, N.G., Rubin-Bejerano, I., Sagee, S., and Shenhar, G. (2003) Transcriptional regulation of meiosis in budding yeast. *Int. Rev. Cytol.*, **224**, 111–171.

47 Johnston, M. (1987) A model fungal gene regulatory mechanism: the *GAL* genes of *Saccharomyces cerevisiae*. *Microbiol. Rev.*, **51** (4), 458–476.

48 Hawkins, K.M. and Smolke, C.D. (2006) The regulatory roles of the galactose permease and kinase in the induction response of the *GAL* network in *Saccharomyces cerevisiae*. *J. Biol. Chem.*, **281** (19), 13 485–13 492.

49 Sikorski, R.S. and Hieter, P. (2002) A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in *Saccharomyces cerevisiae*. *Genetics*, **122** (1), 19–27.

50 Bajwa, W., Torchia, T.E., and Hopper, J.E. (1988) Yeast regulatory gene *GAL3*: carbon regulation; UASGal elements in common with *GAL1*, *GAL2*, *GAL7*, *GAL10*, *GAL80*, and *MEL1*; encoded protein strikingly similar to yeast and *Escherichia coli* galactokinases. *Mol. Cell. Biol.*, **8** (8), 3439–3447.

51 Adams, B.G. (1972) Induction of galactokinase in *Saccharomyces cerevisiae*: kinetics of induction and glucose effects. *J. Bacteriol.*, **111** (2), 308–315.

52 St John, T.P. and Davis, R.W. (1981) The organization and transcription of the galactose gene cluster of *Saccharomyces*. *J. Mol. Biol.*, **152** (2), 285–315.

53 Douglas, H.C. and Hawthorne, D.C. (1964) Enzymatic expression and genetic linkage of genes controlling galactose utilization in *Saccharomyces*. *Genetics*, **49**, 837–844.

54 Johnston, M. and Davis, R.W. (1984) Sequences that regulate the divergent *GAL1-GAL10* promoter in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.*, **4** (8), 1440–1448.

55 West, R.W., Yocum, R.R., and Ptashne, M. (1984) *Saccharomyces cerevisiae GAL1-GAL10* divergent promoter region: location and function of the upstream activating sequence UASG. *Mol. Cell. Biol.*, **4** (11), 2467–2478.

56 Nakao, J., Miyanohara, A., Toh-e, A., and Matsubara, K. (1986) *Saccharomyces cerevisiae PHO5* promoter region: location and function of the upstream activation site. *Mol. Cell. Biol.*, **6** (7), 2613–2623.

57 Bajwa, W., Rudolph, H., and Hinnen, A. (1987) *PHO5* upstream sequences confer phosphate control on the constitutive *PHO3* gene. *Yeast*, **3** (1), 33–42.

58 Miyanohara, A., Toh-e, A., Nozaki, C., Hamada, F., Ohtomo, N., and Matsubara, K. (1983) Expression of hepatitis B surface antigen gene in yeast. *Proc. Natl. Acad. Sci. U.S.A.*, **80** (1), 1–5.

59 Jeppsson, M., Johansson, B., Jensen, P.R., Hahn-Hägerdal, B., and Gorwa-Grauslund, M.F. (2003) The level of glucose-6-phosphate dehydrogenase activity strongly influences xylose fermentation and inhibitor sensitivity in recombinant *Saccharomyces cerevisiae* strains. *Yeast*, **20** (15), 1263–1272.

60 Robinson, A.S., Bockhaus, J.A., Voegler, A.C., and Wittrup, K.D. (1996) Reduction of BiP levels decreases heterologous protein secretion in *Saccharomyces cerevisiae*. *J. Biol. Chem.*, **271** (17), 10 017–10 022.

61 Thiele, D.J. and Hamer, D.H. (1986) Tandemly duplicated upstream control sequences mediate copper-induced transcription of the *Saccharomyces cerevisiae* copper-metallothionein gene. *Mol. Cell. Biol.*, **6** (4), 1158–1163.

62 Fürst, P., Hu, S., Hackett, R., and Hamer, D. (1988) Copper activates metallothionein gene transcription by altering the conformation of a specific DNA binding protein. *Cell*, **55** (4), 705–717.

63 Etcheverry, T. (1990) Induced expression using yeast copper metallothionein promoter. *Methods Enzymol.*, **185**, 319–329.

64 Labbé, S., Zhu, Z., and Thiele, D.J. (1997) Copper-specific transcriptional repression of yeast genes encoding critical components in the copper transport pathway. *J. Biol. Chem.*, **272** (25), 15 951–15 958.

65 Jensen, L.T. and Winge, D.R. (1998) Identification of a copper-induced intramolecular interaction in the transcription factor Mac1 from *Saccharomyces cerevisiae*. *EMBO J.*, **17** (18), 5400–5408.

66 Labbé, S. and Thiele, D.J. (1999) [8] Copper ion inducible and repressible promoter systems in yeast. *Methods Enzymol.*, **306**, 145–153.

67 Abramova, N.E., Cohen, B.D., Sertil, O., Kapoor, R., Davies, K.J., and Lowry, C.V. (2001) Regulatory mechanisms controlling expression of the *DAN/TIR* mannoprotein genes during anaerobic remodeling of the cell wall in *Saccharomyces cerevisiae*. *Genetics*, **157** (3), 1169–1177.

68 Nevoigt, E., Fischer, C., Mucha, O., Matthäus, F., Stahl, U., and Stephanopoulos, G. (2007) Engineering promoter regulation. *Biotechnol. Bioeng.*, **96** (3), 550–558.

69 Gancedo, J.M. (1998) Yeast carbon catabolite repression. *Microbiol. Mol. Biol. Rev.*, **62** (2), 334–361.

70 Weinhandl, K., Winkler, M., Glieder, A., and Camattari, A. (2014) Carbon source dependent promoters in yeasts. *Microb. Cell Fact.*, **13** (1), 5.

71 Denis, C.L. and Young, E.T. (1983) Isolation and characterization of the positive regulatory gene *ADR1* from *Saccharomyces cerevisiae*. *Mol. Cell. Biol.*, **3** (3), 360–370.

72 Shuster, J., Yu, J., Cox, D., Chan, R.V., Smith, M., and Young, E. (1986) ADR1-mediated regulation of *ADH2* requires an inverted repeat sequence. *Mol. Cell. Biol.*, **6** (6), 1894–1902.

73 Price, V.L., Taylor, W.E., Clevenger, W., Worthington, M., and Young, E.T. (1990) Expression of heterologous proteins in *Saccharomyces cerevisiae* using the *ADH2* promoter. *Methods Enzymol.*, **185**, 308–318.

74 Chambers, P., Issaka, A., and Palecek, S.P. (2004) *Saccharomyces cerevisiae JEN1* promoter activity is inversely related to concentration of repressing sugar. *Appl. Environ. Microbiol.*, **70** (1), 8–17.

75 Da Silva, N.A. and Srikrishnan, S. (2012) Introduction and expression of genes for metabolic engineering applications in *Saccharomyces cerevisiae*. *FEMS Yeast Res.*, **12** (2), 197–214.

76 Peng, B., Williams, T.C., Henry, M., Nielsen, L.K., and Vickers, C.E. (2015) Controlling heterologous gene expression in yeast cell factories on different carbon substrates and across the diauxic shift: a comparison of yeast promoter activities. *Microb. Cell Fact.*, **14** (1), 91.

77 Holland, M.J. and Holland, J.P. (1978) Isolation and identification of yeast messenger ribonucleic acids coding for enolase, glyceraldehyde-3-phosphate dehydrogenase, and phosphoglycerate kinase. *Biochemistry*, **17** (23), 4900–4907.

78 Hommes, F.A. (1966) Effect of glucose on the level of glycolytic enzymes in yeast. *Arch. Biochem. Biophys.*, **114** (1), 231–233.

79 Denis, C.L., Ferguson, J., and Young, E.T. (1983) mRNA levels for the fermentative alcohol dehydrogenase of *Saccharomyces cerevisiae* decrease upon growth on a nonfermentable carbon source. *J. Biol. Chem.*, **258** (2), 1165–1171.

80 Partow, S., Siewers, V., Bjørn, S., Nielsen, J., and Maury, J. (2010) Characterization of different promoters for designing a new expression vector in *Saccharomyces cerevisiae*. *Yeast*, **27** (11), 955–964.

81 Chambers, A., Tsang, J.S., Stanway, C., Kingsman, A.J., and Kingsman, S.M. (1989) Transcriptional control of the *Saccharomyces cerevisiae PGK* gene by RAP1. *Mol. Cell. Biol.*, **9** (12), 5516–5524.

82 Baker, H.V. (1991) *GCR1* of *Saccharomyces cerevisiae* encodes a DNA binding protein whose binding is abolished by mutations in the CTTCC sequence motif. *Proc. Natl. Acad. Sci. U.S.A.*, **88** (21), 9443–9447.

83 Bitter, G.A., Chang, K.K., and Egan, K.M. (1991) A multi-component upstream activation sequence of the *Saccharomyces cerevisiae* glyceraldehyde-3-phosphate dehydrogenase gene promoter. *Mol. Gen. Genet.*, **231** (1), 22–32.

84 Holland, M.J., Yokoi, T., Holland, J.P., Myambo, K., and Innis, M.A. (1987) The *GCR1* gene encodes a positive transcriptional regulator of the enolase and glyceraldehyde-3-phosphate dehydrogenase gene families in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.*, **7** (2), 813–820.

85 Pfeifer, K., Kim, K.S., Kogan, S., and Guarente, L. (1989) Functional dissection and sequence of yeast HAP1 activator. *Cell*, **56** (2), 291–301.

86 Munholland, J.M., Kelly, J.K., and Wildeman, A.G. (1990) DNA sequences required for yeast actin gene transcription do not include conserved CCAAT motifs. *Nucleic Acids Res.*, **18** (20), 6061–6068.

87 Cottrelle, P., Thiele, D., Price, V.L., Memet, S., Micouin, J.Y., Marck, C., Buhler, J.M., Sentenac, A., and Fromageot, P. (1985) Cloning, nucleotide sequence, and expression of one of two genes coding for yeast elongation factor 1 alpha. *J. Biol. Chem.*, **260** (5), 3090–3096.

88 Jensen, P.R. and Hammer, K. (1998) Artificial promoters for metabolic optimization. *Biotechnol. Bioeng.*, **58** (2-3), 191–195.

89 Blount, B.A., Weenink, T., Vasylechko, S., and Ellis, T. (2012) Rational diversification of a promoter providing fine-tuned expression and orthogonal regulation for synthetic biology. *PLoS One*, **7** (3), e33 279.

90 Blazeck, J., Garg, R., Reed, B., and Alper, H.S. (2012) Controlling promoter strength and regulation in *Saccharomyces cerevisiae* using synthetic hybrid promoters. *Biotechnol. Bioeng.*, **109** (11), 2884–2895.

91 Blazeck, J. and Alper, H.S. (2013) Promoter engineering: recent advances in controlling transcription at the most fundamental level. *Biotechnol. J.*, **8** (1), 46–58.

92 Alper, H., Fischer, C., Nevoigt, E., and Stephanopoulos, G. (2005) Tuning genetic control through promoter engineering. *Proc. Natl. Acad. Sci. U.S.A.*, **102** (36), 12 678–12 683.

93 Nevoigt, E., Kohnke, J., Fischer, C.R., Alper, H., Stahl, U., and Stephanopoulos, G. (2006) Engineering of promoter replacement cassettes for fine-tuning of gene expression in *Saccharomyces cerevisiae*. *Appl. Environ. Microbiol.*, **72** (8), 5266–5273.

94 Blake, W.J., Balázsi, G., Kohanski, M.A., Isaacs, F.J., Murphy, K.F., Kuang, Y., Cantor, C.R., Walt, D.R., and Collins, J.J. (2006) Phenotypic consequences of promoter-mediated transcriptional noise. *Mol. Cell*, **24** (6), 853–865.

95 Levo, M., Zalckvar, E., Sharon, E., Dantas Machado, A.C., Kalma, Y., Lotam-Pompan, M., Weinberger, A., Yakhini, Z., Rohs, R., and Segal, E. (2015) Unraveling determinants of transcription factor binding outside the core binding site. *Genome Res.*, **25** (7), 1018–1029.

96 Curran, K.A., Crook, N.C., Karim, A.S., Gupta, A., Wagman, A.M., and Alper, H.S. (2014) Design of synthetic yeast promoters via tuning of nucleosome architecture. *Nat. Commun.*, **5**, Article number 4002. doi: 10.1038/ncomms5002.

97 Garí, E., Piedrafita, L., Aldea, M., and Herrero, E. (1997) A set of vectors with a tetracycline-regulatable promoter system for modulated gene expression in *Saccharomyces cerevisiae*. *Yeast*, **13** (9), 837–848.

98 Redden, H. and Alper, H.S. (2015) The development and characterization of synthetic minimal yeast promoters. *Nat. Commun.*, **6**, 7810.

99 Khalil, A.S., Lu, T.K., Bashor, C.J., Ramirez, C.L., Pyenson, N.C., Joung, J.K., and Collins, J.J. (2012) A synthetic biology framework for programming eukaryotic transcription functions. *Cell*, **150** (3), 647–658.

100 McIsaac, R.S., Oakes, B.L., Wang, X., Dummit, K.A., Botstein, D., and Noyes, M.B. (2013) Synthetic gene expression perturbation systems with rapid, tunable, single-gene specificity in yeast. *Nucleic Acids Res.*, **41** (4), e57.

101 Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.*, **30** (6), 521–530.

102 Ottoz, D.S., Rudolf, F., and Stelling, J. (2014) Inducible, tightly regulated and growth condition-independent transcription factor in *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **42** (17), e130.

103 Murphy, K.F., Balázsi, G., and Collins, J.J. (2007) Combinatorial promoter design for engineering noisy gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **104** (31), 12 726–12 731.

104 Estojak, J., Brent, R., and Golemis, E.A. (1995) Correlation of two-hybrid affinity data with *in vitro* measurements. *Mol. Cell. Biol.*, **15** (10), 5820–5829.

105 Ellis, T., Wang, X., and Collins, J.J. (2009) Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nat. Biotechnol.*, **27** (5), 465–471.

106 McIsaac, R.S., Gibney, P.A., Chandran, S.S., Benjamin, K.R., and Bostein, D. (2014) Synthetic biology tools for programming gene expression without nutritional perturbations in *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **42** (6), e68.

107 Brent, R. and Ptashne, M. (1984) A bacterial repressor protein or a yeast transcriptional terminator can block upstream activation of a yeast gene. *Nature*, **312** (5995), 612–615.

108 Golemis, E.A. and Brent, R. (1992) Fused protein domains inhibit DNA binding by LexA. *Mol. Cell. Biol.*, **12** (7), 3006–3014.

109 Ruden, D.M., Ma, J., Li, Y., Wood, K., and Ptashne, M. (1991) Generating yeast transcriptional activators containing no yeast protein sequences. *Nature*, **350** (6315), 250–252.

110 Dingermann, T., Frank-Stoll, U., Werner, H., Wissmann, A., Hillen, W., Jacquet, M., and Marschalek, R. (1992) RNA polymerase III catalysed transcription can be regulated in *Saccharomyces cerevisiae* by the bacterial tetracycline repressor-operator system. *EMBO J.*, **11** (4), 1487–1492.

111 Nagahashi, S., Nakayama, H., Hamada, K., Yang, H., Arisawa, M., and Kitada, K. (1997) Regulation by tetracycline of gene expression in *Saccharomyces cerevisiae*. *Mol. Gen. Genet.*, **255** (4), 372–375.

112 Mnaimneh, S., Davierwala, A.P., Haynes, J., Moffat, J., Peng, W.T., Zhang, W., Yang, X., Pootoolal, J., Chua, G., Lopez, A., Trochesset, M., Morse, D., Krogan, N.J., Hiley, S.L., Li, Z., Morris, Q., Grigull, J., Mitsakakis, N., Roberts, C.J., Greenblatt, J.F., Boone, C., Kaiser, C.A., Andrews, B.J., and Hughes, T.R. (2004) Exploration of essential gene functions via titratable promoter alleles. *Cell*, **118** (1), 31–44.

113 Zhang, N., Osborn, M., Gitsham, P., Yen, K., Miller, J.R., and Oliver, S.G. (2003) Using yeast to place human genes in functional categories. *Gene*, **303**, 121–129.

114 Bellí, G., Garí, E., Piedrafita, L., Aldea, M., and Herrero, E. (1998) An activator/repressor dual system allows tight tetracycline-regulated gene expression in budding yeast. *Nucleic Acids Res.*, **26** (4), 942–947.

115 Klug, A. (2010) The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Annu. Rev. Biochem.*, **79** (1), 213–231.

116 Garg, A., Lohmueller, J.J., Silver, P.A., and Armel, T.Z. (2012) Engineering synthetic TAL effectors with orthogonal target sites. *Nucleic Acids Res.*, **40** (15), 7584–7595.

117 Gilbert, L.A., Larson, M.H., Morsut, L., Liu, Z., Brar, G.A., Torres, S.E., Stern-Ginossar, N., Brandman, O., Whitehead, E.H., Doudna, J.A., Lim, W.A., Weissman, J.S., and Qi, L.S. (2013) CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell*, **154** (2), 442–451.

118 Farzadfard, F., Perli, S.D., and Lu, T.K. (2013) Tunable and multifunctional eukaryotic transcription factors based on CRISPR/Cas. *ACS Synth. Biol.*, **2** (10), 604–613.

119 Koleske, A.J. and Young, R.A. (1994) An RNA polymerase II holoenzyme responsive to activators. *Nature*, **368** (6470), 466–469.

120 Barberis, A., Pearlberg, J., Simkovich, N., Farrell, S., Reinagel, P., Bamdad, C., Sigal, G., and Ptashne, M. (1995) Contact with a component of the polymerase II holoenzyme suffices for gene activation. *Cell*, **81** (3), 359–368.

121 Gill, G. and Ptashne, M. (1987) Mutants of GAL4 protein altered in an activation function. *Cell*, **51** (1), 121–126.

122 Hope, I.A., Mahadevan, S., and Struhl, K. (1988) Structural and functional characterization of the short acidic transcriptional activation region of yeast GCN4 protein. *Nature*, **333** (6174), 635–640.

123 Ma, J. and Ptashne, M. (1987) A new class of yeast transcriptional activators. *Cell*, **51** (1), 113–119.

124 Fields, S. and Song, O. (1989) A novel genetic system to detect protein-protein interactions. *Nature*, **340** (6230), 245–246.

125 Gyuris, J., Golemis, E., Chertkov, H., and Brent, R. (1993) Cdi1, a human G1 and S phase protein phosphatase that associates with Cdk2. *Cell*, **75** (4), 791–803.

126 Shimizu-Sato, S., Huq, E., Tepperman, J.M., and Quail, P.H. (2002) A light-switchable gene promoter system. *Nat. Biotechnol.*, **20** (10), 1041–1044.

127 McIsaac, R.S., Silverman, S.J., McClean, M.N., Gibney, P.A., Macinskas, J., Hickman, M.J., Petti, A.A., and Botstein, D. (2011) Fast-acting and nearly gratuitous induction of gene expression and protein depletion in *Saccharomyces cerevisiae*. *Mol. Biol. Cell*, **22** (22), 4447–4459.

128 Louvion, J.F., Havaux-Copf, B., and Picard, D. (1993) Fusion of GAL4-VP16 to a steroid-binding domain provides a tool for gratuitous induction of galactose-responsive genes in yeast. *Gene*, **131** (1), 129–134.

129 Brambilla, A., Mainieri, D., and Agostoni Carbone, M.L. (1997) A simple signal element mediates transcription termination and mRNA 3' end formation in the *DEG1* gene of *Saccharomyces cerevisiae*. *Mol. Gen. Genet.*, **254** (6), 681–688.

130 Wang, M., Li, S., and Zhao, H. (2016) Design end engineering of intracellular-metabolite-sensing/regulation gene circuits in *Saccharomyces cerevisiae*. *Biotechnol. Bioeng.*, **113** (1), 206–215.

131 Nevozhay, D., Adams, R.M., Murphy, K.F., Josic, K., and Balázsi, G. (2009) Negative autoregulation linearizes the dose-response and suppresses the heteogeneity of gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **106** (13), 5123–5128.

132 Nevozhay, D., Zal, T., and Balázsi, G. (2013) Transferring a synthetic gene circuit from yeast to mammalian cells. *Nat. Commun.*, **4**, 1451.

133 Keleher, C.A., Redd, M.J., Schultz, J., Carlson, M., and Johnson, A.D. (1992) Ssn6-Tup1 is a general repressor of transcription in yeast. *Cell*, **68** (4), 709–719.

134 Tzamarias, D. and Struhl, K. (1994) Functional dissection of the yeast Cyc8-Tup1 transcriptional co-repressor complex. *Nature*, **369** (6483), 758–761.

135 Nevoigt, E. (2008) Progress in metabolic engineering of *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.*, **72** (3), 379–412.

136 Blount, B.A., Weenink, T., and Ellis, T. (2012) Construction of synthetic regulatory networks in yeast. *FEBS Lett.*, **586** (15), 2112–2121.

137 Redden, H., Morse, N., and Alper, H.S. (2015) The synthetic biology toolbox for tuning gene expression in yeast. *FEMS Yeast Res.*, **15** (1), 1–10.

138 Endy, D. (2005) Foundations for engineering biology. *Nature*, **438** (7067), 449–453.

139 La Russa, M.F. and Qi, L.S. (2015) The new state of the art: Cas9 for gene activation and repression. *Mol. Cell. Biol.*, **35** (22), 3800–3809.

140 Reider Apel, A.A., d'Espaux, L., Wehrs, M., Sachs, D., Li, R.A., Tong, G.J., Garber, M., Nnadi, O., Zhuang, W., Hillson, N.J., Keasling, J.D., and Mukhopadhyay, A. (2017) A Cas9-based toolkit to program gene expression in *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **45** (1), 496–508.

7

Splicing and Alternative Splicing Impact on Gene Design

Beatrix Suess, Katrin Kemmerer, and Julia E. Weigand

Department of Biology, Technische Universität Darmstadt, Schnittspahnstraße 10, 64287 Darmstadt, Germany

Most human genes are interrupted by one or more introns that have to be removed to generate mRNAs with intact open reading frames (ORFs), a process called pre-mRNA splicing. A ribonucleoprotein complex, the spliceosome, is responsible for the accurate removal of the intervening sequences. Alternative splicing, that is, not all exons are included in the mature mRNA every time, creates the possibility that one gene can encode for more than one protein. This immensely increases the coding capacity of a genome. In humans, aberrant splicing has been recognized to be the causative agent of several hereditary diseases and to drive cancer progression. In contrast to humans, introns are rare in budding yeast but seem to be important for fine-tuning gene expression and growth under stress conditions.

7.1 The Discovery of "Split Genes"

In 1977, Richard J. Roberts and Phillip A. Sharp studied adenovirus type 2, a double-stranded DNA virus causing common cold. Their aim was to map the location of the genes on the viral genome. Unexpectedly, they found that the mRNA did not hybridize to the DNA in a continuous stretch. Instead, it hybridized to four neighboring segments in the genome, separated by three intervening sequences. These intervening sequences were looped out in the DNA as they were missing in the mRNA sequence [1, 2]. This came as a surprise, as former analyses of bacterial genes suggested that a gene comprises a continuous stretch of DNA. Soon after this initial discovery, this discontinuous gene structure was shown to be a common feature of eukaryotic genes. Sixteen years later, both were awarded the Nobel Prize in Physiology or Medicine for their discovery of "split genes."

The realization that eukaryotic genes are comprised of exons (sequences of a gene included into the mature mRNA) and introns (intervening sequences removed upon splicing) called for a new mRNA maturation process: the removal of the intronic sequences from the pre-mRNA to yield a shortened mature

Synthetic Biology: Parts, Devices and Applications, First Edition. Edited by Christina Smolke.

mRNA (splicing). Subsequently a complex machinery that deletes the intervening sequences of the pre-mRNA was identified: the spliceosome. The discovery that not all exons are included in the mature mRNA every time came as a further surprise. This process was appropriately called alternative splicing and opened up the possibility that one gene could code for more than one protein.

Alternative splicing is highly regulated during development and different mRNA isoforms are important for determining the fate of different cell types and tissues. Therefore, (alternative) splicing is viewed as an integral part of mRNA maturation in eukaryotes, and aberrant splicing has not only been recognized to be the causative agent of several hereditary diseases but also to drive cancer progression.

7.2 Nuclear Pre-mRNA Splicing in Mammals

7.2.1 Introns and Exons: A Definition

The average human gene contains eight exons with a mean length of 145 nucleotides and introns more than ten times this size [3]. *Cis*-acting elements encoded in the pre-mRNA provide the information that defines an intron (see Figure 7.1). The 5′ splice site marks the beginning of the intron and includes the dinucleotide GU encompassed within a larger, less conserved consensus sequence. The 3′ end of the intron carries three conserved sequence elements. The branch point is usually an adenosine located within a less conserved sequence element (branch site), typically located 18–40 nucleotides upstream from the 3′ splice site. It is followed by the polypyrimidine tract and a terminal AG dinucleotide at the extreme 3′ end of the intron [4, 5]. The vast majority of introns contain the canonical splice sites GU-AG (99%). However, other categories exist that occur rarely, including the noncanonical splice sites GC-AG and AU-AC [6].

7.2.2 The Catalytic Mechanism of Splicing

The splicing process consists of two consecutive transesterification reactions. In the first step, the 5′ exon–intron junction is attacked by a free hydroxyl group provided internally by the 2′ hydroxyl group from the branch point adenosine. This leads to cleavage at the 5′ splice site and ligation of the 5′ end of the intron to the 2′ hydroxyl group of the branch point adenosine. In the second step, the free 3′ hydroxyl group of the released 5′ exon in turn attacks the phosphate at the 3′ intron–exon border. This results in ligation of the two exons and the release of the intron in form of a lariat (reviewed in [5, 7, 8]).

7.2.3 A Complex Machinery to Remove Nuclear Introns: The Spliceosome

Splicing is catalyzed by the spliceosome, a large and highly dynamic macromolecular ribonucleoprotein complex that assembles on the intron-containing pre-mRNA. The major spliceosome consists of the U1, U2, U4/U6, and U5 small

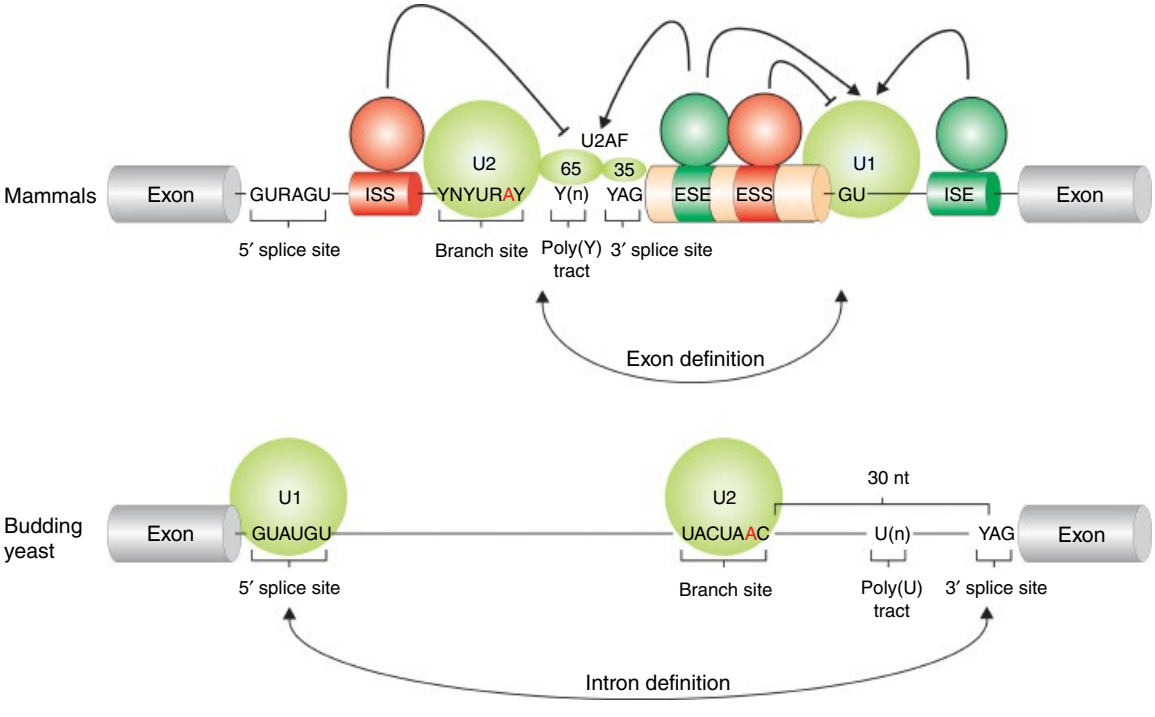


Figure 7.1 Conserved sequence elements of mammalian and budding yeast pre-mRNAs. Exons (cylinders) are separated by introns (lines). The consensus sequences in mammals and budding yeast at the 5′ splice site, branch site, and 3′ splice site are as indicated. N is any nucleotide, R is purine, and Y is pyrimidine. Mammals contain a polypyrimidine-rich stretch; *S. cerevisiae* contains a polyuridine-rich stretch. Both are located between the branch site and the 3′ splice site. In mammals, cross-exon complexes are formed during early stages of spliceosome assembly, while in *S. cerevisiae* the introns are defined. The spliceosomal snRNPs U1 and U2 (green) are shown interacting with the splice sites. Mammals additionally have the U2 auxiliary factor (U2AF), U2AF65 and U2AF35 (green) interacting with the 3′ splice site. They also use auxiliary regulatory elements that either enhance the splicing process, namely, exonic and intronic splicing enhancers (ESE and ISE, dark green cylinders), or inhibit spliceosome assembly, such as exonic and intronic splicing silencers (ESS and ISS, red cylinders). These elements are often bound by SR proteins and hnRNPs.

nuclear ribonucleoprotein particles (snRNPs). Each snRNP consists of an snRNA (two in the case of U4/U6) and seven Sm proteins that form a ring-shaped structure (U6, as an exception, contains Sm-like proteins). Each snRNP contains additionally a variable number of particle-specific proteins. Furthermore, a large number of auxiliary proteins assemble co-transcriptionally on nascent pre-mRNAs to accurately recognize the splice sites [5, 9–11].

The *cis*-acting pre-mRNA sequence elements help to define the splice sites and mediate interactions between the pre-mRNA and components of the spliceosome [12–14]. The 5′ splice site interacts with the U1 snRNP via base pairing between the splice site and the 5′ end of the U1 snRNA. The 3′ end is consecutively recognized by several proteins, including non-snRNP factors like splicing factor 1 (SF1), which binds to the branch point. The U2 auxiliary factor (U2AF), a heterodimer consisting of a 65 and a 35 kDa subunit, binds the polypyrimidine tract and the 3′ splice site. These factors form the early (E) complex. In a subsequent step, the E complex is joined by the U2 snRNP that binds to the branch point forming the A complex. This structure is then bound by the preassembled tri-snRNP consisting of the U5 and the U4/U6 snRNPs, generating the precatalytic B complex. The B complex undergoes major rearrangements in RNA–RNA and RNA–protein interactions, leading to the destabilization of U1 and U4 snRNP binding. This catalytically activates the B complex to mediate the first catalytic step of splicing and yields the C complex, which in turn catalyzes the second step. The spliceosome then dissociates and is recycled for additional rounds of splicing [5, 7, 11]. Lately several high-resolution structures of different spliceosomal complexes from budding yeast and humans have been solved using cryo-electron microscopy (e.g., [15–17]). These structures give unprecedented insight on the architecture of the different complexes and aid our understanding of the structural rearrangements that have to occur to complete one catalytic cycle.

7.2.4 Exon Definition

When the length of an intron exceeds 200–250 nucleotides, which is the case for most introns in higher eukaryotes, early splicing complexes form across an exon [18], a process called exon definition [19]. During exon definition, the U1 snRNP binds to the 5′ splice site downstream of an exon and promotes the association of U2AF with the polypyrimidine tract at the upstream 3′ splice site. This leads subsequently to the recruitment of the U2 snRNP to the branch point upstream of the exon. The complex is stabilized by the binding of additional proteins of the serine/arginine (SR) protein family (see Section 7.5.2) to enhancer elements within the exon [20, 21]. In addition, exon definition might be facilitated by pairs of intronic enhancer elements flanking constitutive as well as alternatively spliced exons [22].

Before proceeding to the splicing reaction, exon-defined complexes must be converted to intron-defined complexes. This requires disruption of the cross-exon interactions, followed by conversion into a cross-intron A complex, in which a molecular bridge is formed from U2 to U1 bound to an upstream 5′ splice site [21, 23]. In an alternative assembly pathway, the tri-snRNP is already

present in the exon definition complex, interacting with the U2 snRNP by base pairing between the U2 and U6 snRNAs [24]. Such complexes can then be directly converted to precatalytic B-like complexes, without prior formation of the cross-intron A complex [24].

Splicing mainly occurs co-transcriptional (see Section 7.5.4) with a 5′ to 3′ directionality. Exceptions to this rule include introns flanking alternatively spliced exons with the excision being delayed or even happening posttranscriptionally [25, 26].

7.3 Splicing in Yeast

7.3.1 Organization and Distribution of Yeast Introns

From an evolutionary perspective, yeasts are a highly diverse group of single-celled microorganisms within the kingdom of fungi. The budding yeast (also "true yeast"), including the well-known *Saccharomyces cerevisiae*, belongs to the phylum Ascomycota. Other yeasts, like the fission yeast *Schizosaccharomyces pombe*, belong to the phylum Basidiomycota. Both *S. cerevisiae* and *S. pombe* are eukaryotic model organisms.

The genome of *S. pombe* was sequenced in 2002 [27]. It contains ~4800 genes with ~43% of the genes containing up to 15 introns. The average intron size is 81 nucleotides. With regard to splicing factors and 3′ splice site selection, splicing in *S. pombe* is considered to be more similar mechanistically to mammals than in *S. cerevisiae* [28, 29]. We will still focus exclusively on budding yeast in the following sections, as *S. cerevisiae* is the more widely studied eukaryotic model organism.

In contrast to mammals (and fission yeast), very few genes in budding yeast code for introns. Only 5% of the ~5800 genes contain introns [30]. In addition, most genes contain only one intronic sequence; a mere 10 genes code for 2 introns. Why does *S. cerevisiae* have such an intron-poor genome? In general, unicellular eukaryotes seem to be under pressure to loose introns. A correlation exists between the intron density of a genome and the logarithm of the generation time of an organism: organisms with a short generation time tend to have fewer introns when compared with more slowly growing organisms [31]. This observation could be explained by selection for smaller genomes and for faster protein production, for example, in response to stress conditions.

Intron boundaries in *S. cerevisiae* are well defined, with a 6 bp sequence at the 5′ splice site and a 7 bp sequence at the branch site required for efficient splicing (see Figure 7.1) [32–34]. The average distance between the branch point and 3′ splice site is 30 nucleotides and this region also contains a poly(U) tract (see Figure 7.1) [30, 35]. Introns tend to be short, with an average length of 154 nucleotides in non-ribosomal and 408 nucleotides in ribosomal proteins.

Introns are not equally distributed in the *S. cerevisiae* genome, but are highly enriched in ribosomal proteins. Eighty nine of the 137 ribosomal proteins (>60%) code for at least one intron, whereas only 198 of the remaining genes (<1%)

contain introns [30]. Furthermore, the location of introns within a gene is strongly biased toward its 5′ end [31]. This bias is thought to arise due to homologous recombination of a gene's cDNA with its genomic copy and could simultaneously explain how introns might have been lost during yeast evolution. cDNA arises by reverse transcription of mRNA, which does not contain introns and is a by-product of the activity of retrotransposons. Reverse transcription starts at the 3′ of the mRNA and often terminates prematurely, which leads to a 3′ bias in cDNAs and, as a result, to preferential loss of introns at the 3′ end of genes after recombination of the cDNA with the genomic copy.

Furthermore, *S. cerevisiae* introns tend to be located in highly expressed mRNAs. 27% of all mRNAs produced per hour are generated from the 5% intron-containing genes [36]. Genome-wide analyses of mRNA [37] and protein [38] levels showed that, on average, intron-containing genes produce ~3.9-fold more RNA and 3.3-fold more protein than intronless genes.

7.4 Splicing without the Spliceosome

7.4.1 Group I and Group II Self-Splicing Introns

Interrupted genes are found not only in the genomes of yeast and metazoan, but are present in all classes of organisms. The majority of introns are spliced out by the spliceosome (nuclear pre-mRNA introns). Besides this, self-splicing introns (group I and group II) exist, in which the intervening sequences can excise themselves from the RNA in an autocatalytic manner [39].

Group I and II introns are found in the DNA of organelles, bacteria, and the nucleus of lower eukaryotes (group I only). Their occurrence is more sporadic in bacteria than in lower eukaryotes and is most common in the organelles of higher plants. Whereas group II introns are mainly found in organelles, group I introns interrupt rRNA, mRNA, and tRNA in bacteria, as well as in the organelles of lower eukaryotes, and some plants. In addition, they have been found in several bacteriophages.

Nuclear pre-mRNA introns are defined by *cis*-acting sequence elements that are recognized by the spliceosome. Group I and group II introns, in contrast, adopt a typical secondary structure that contains distinct domains, which then folds into a highly complex tertiary structure. As a consequence, the catalytic mechanism of this splicing reaction solely depends on the sequence and the correct folding of the intron. The RNA tertiary structure brings the 5′ and the 3′ splice sites in close proximity and generates a catalytic site. The fold is stabilized by several magnesium ions, allowing the RNA to perform the splicing reaction *in vitro* by itself, without any enzymatic activities provided by proteins. Proteins are required only to assist correct folding of the complex structure *in vivo*. For this fundamental discovery that RNA can harbor catalytic function, Tom R. Cech (together with Sidney Altman) was awarded with the Nobel Prize in Chemistry in 1989 [40].

For group I introns, the only factors required for autosplicing are monovalent and divalent cations and a guanine nucleotide cofactor. The 3′ hydroxyl group of

the cofactor attacks the 5′ end of the intron, resulting in the first transesterification reaction. The free 3′ hydroxyl of the first exon thus generated then attacks the junction between intron and second exon, leading to the second transesterification step. Consequently, the intron is released as a linear molecule that circularizes later [41].

Group II introns share the same catalytic mechanism as nuclear pre-mRNA introns excised by the spliceosome with the first nucleophilic attack of a branch point adenosine, resulting in lariat formation of the intron (see Section 7.2.2) [42]. Interestingly, recent data indicate that the U6 snRNA of the spliceosome catalyzes both splicing steps by positioning divalent metal ions so that they stabilize the leaving group during each reaction. Notably, all ligands of the catalytically active metal ions in the U6 snRNA correspond to ligands observed to position catalytically active divalent metals in the crystal structures of group II intron RNAs [43]. This agreement indicates that group II introns and the spliceosome share common catalytic mechanisms and probably common evolutionary origins [44]. It also suggests that splicing evolved from an autocatalytic reaction inherent to an individual RNA molecule [45]. As splicing became more complex, proteins started to play a more important role. Importantly, the similarities between the catalytic core of the group II intron and the U6 snRNA support the hypothesis that spliceosomal introns in eukaryotes developed out of group II self-splicing introns [46].

7.4.2 tRNA Splicing

The splicing of tRNAs in archaea and eukarya is the only example of intron removal that does not involve transesterification, but instead successive cleavage and ligation reactions. tRNAs contain a single intron located one nucleotide next to the anticodon. These introns are short (14–60 nucleotides) and have no consensus sequence. They are recognized by an endonuclease that detects a common secondary structure of the tRNA rather than a sequence element. It cleaves both ends of the intron generating two tRNA halves that are subsequently joined by an RNA ligase [47].

7.5 Alternative Splicing in Mammals

7.5.1 Different Mechanisms of Alternative Splicing

Alternative splicing affects 95% of all human genes [48, 49] and produces multiple mRNA molecules from a single gene. The resulting proteomic diversity is important for many different cellular processes, including cell growth and differentiation [13].

Alternative splicing events can be divided into four major categories: inclusion and exclusion of (cassette) exons, the usage of alternative 5′ or 3′ splice sites, and the retention of entire introns (see Figure 7.2). Of these, the cassette exon type accounts for approximately one third of all alternative splicing events in humans [50]. Cassette exons are either fully included or excluded in the mature mRNA.

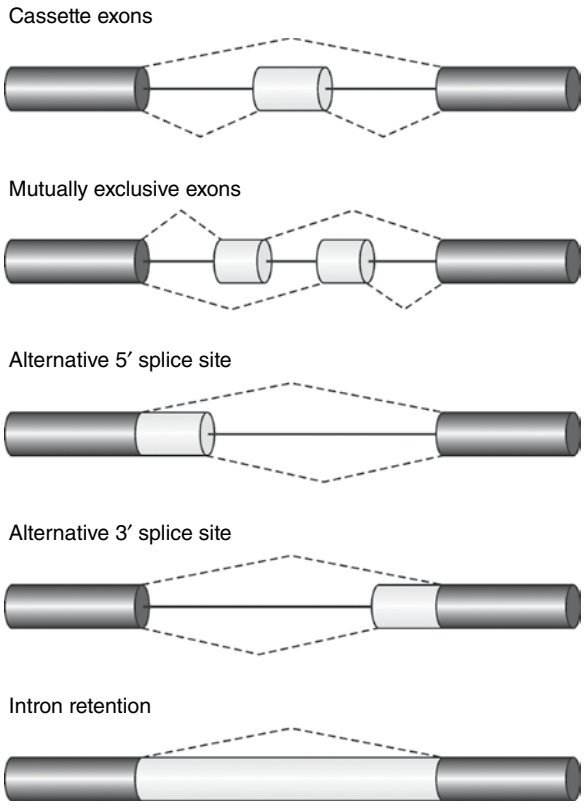


Figure 7.2 Alternative splicing events in mammalian transcripts. The main types of alternative splicing, which are responsible for the generation of different transcripts, are depicted. Dark gray indicate constitutive, and light gray cylinders alternative exons.

In certain cases, multiple cassette exons can be mutually exclusive, producing mRNAs that always include one of several possible exon choices, but not more. Additionally, the use of alternative 5′ or 3′ splice sites can lengthen or shorten exons, a mechanism that accounts for 25% of all alternative splicing events [50]. Finally, the failure to remove an intron leads to a splicing pattern called intron retention. All four types can occur in the translated or untranslated regions (UTRs) of any given pre-mRNA [51].

Many genes show multiple splicing patterns, often in conjunction with the usage of alternative promoters or polyadenylation sites. One striking example is the fast skeletal troponin T (*tnnt3*) gene, which is part of the troponin complex and undergoes extensive alternative splicing. The *tnnt3* gene encodes 19 exons, including five alternatively spliced exons (exons 4–8) and a pair of mutually exclusive exons (exons 16 and 17) [52]. While isoforms including exon 17 (or β) are predominantly expressed throughout development, exon 16 (or α)-containing isoforms are mostly abundant in adult muscles [53, 54]. In addition, the *tnnt3* gene contains a developmentally regulated fetal exon F located between exons 8 and 9 [55, 56].

Hence, multiple isoforms with structural differences are produced and regulated during muscle development and adaptation. Changes in splicing enable transitions from large to small, from acidic to basic isoforms during muscle development [57]. Furthermore, the resulting protein isoforms show differences in their sensitivity to Ca^{2+} activation and their cooperativity of contraction [58–62]. Recently, differential expression patterns of *tnnt3* pre-mRNAs were observed in rat skeletal muscle in response to variation in body weight and also in C2C12 muscle cells upon mechanical stretching [63, 64].

Apart from physiological adaptations, aberrant splicing of the *tnnt3* gene may contribute to disease development. An aberrant splicing pattern was identified in myotonic diseases type 1 and 2 [65]. In mice overexpressing FRG1 (FSHD region gene 1), aberrant splicing of the *tnnt3* pre-mRNA leads to an anomalous fast skeletal troponin T isoform that characterizes dystrophic symptoms [66].

7.5.2 Auxiliary Regulatory Elements

To allow for a correct decision as to which exon is removed or included, additional RNA sequence elements and regulatory proteins are required. A genome-wide study of alternative splicing in mammalian tissues revealed an important role of RNA-binding proteins in splicing regulation via their interaction with *cis*-acting regulatory elements [67]. The relevant RNA sequence elements are categorized depending on their function and position. Sequences enhancing the splicing reaction are known as exonic splicing enhancer (ESE) or intronic splicing enhancer (ISE), while sequences that inhibit splicing are called exonic splicing silencer (ESS) or intronic splicing silencer (ISS) [9].

In general, splicing regulators appear to exhibit position-dependent effects on splicing outcomes [68–70]. One family of RNA-binding proteins are the SR-rich proteins. To interact with the RNA, they contain one or two N-terminal RNA recognition motifs (RRMs). Additionally, they contain a unique, variable-length RS domain at their carboxyl-terminus that functions as a protein interaction domain [71–73]. The core SR protein family consists of 12 members, named serine/arginine-rich splicing factors SRSF1-SRSF12, respectively [74]. SRSF1 and SRSF2 were discovered for their essential roles in constitutive and alternative splicing [75]. They promote both U1 snRNP binding to the 5′ splice site and U2 snRNP binding to the 3′ splice site, allowing for communication between these recognition events [76–79], facilitating exon definition (see Section 7.2.4). The role of SR proteins in splice site selection is discussed in Section 7.5.3. In addition, individual SR protein expression is subject to extensive auto- and cross regulation [80, 81]. They also interact with chromatin [82], couple with the transcription machinery [83, 84], and are involved in mRNA export [85]. The regulation of SR protein activity occurs at the posttranslational level. Site- or region-specific phosphorylation, catalyzed by specific SR protein kinases, is essential to modulate their functions during different stages of RNA processing (reviewed in [86]).

Another family of splicing regulators is the extended family of heterogeneous nuclear ribonucleoproteins (hnRNPs). This family includes an initially identified set of more than 20 polypeptides, designated hnRNP A to U [87]. Their number has further increased as many splicing isoforms, paralogs, and newly identified

proteins have been added based on structural and functional considerations (reviewed in [88]). Affiliated are the following proteins: Nova, Sam68, QKI, TDP-43, TIA, Hu, Fox, CUG-BP, MBNL, and ESRP proteins [88]. hnRNPs share a modular structure, most frequently containing one or more RRMs, one of the most abundant protein domains found in eukaryotes [89, 90]. The K homology (KH)-type RNA-binding domain occurs in the hnRNP proteins K and E and in the hnRNP-like proteins Nova, Sam68, and QKI [88]. In addition to that, many hnRNPs contain RGG boxes (repeats of Arg-Gly-Gly) and other auxiliary domains, such as acidic and glycine- or proline-rich domains [91, 92]. Most hnRNPs shuttle between nucleus and cytoplasm [93].

A few examples shall give an overview of the multifunctionality of hnRNPs: hnRNP A1 is one of the most abundant and ubiquitously expressed members (reviewed in [94]). Its role is not limited to splicing regulation, but includes functions in transcription [95–97], mRNA stability [98, 99], mRNA export [100], translation [101, 102], and telomere maintenance [103, 104]. Another example is polypyrimidine-tract-binding protein (PTB) or hnRNP I (reviewed in [105]), which is involved in splicing [106], mRNA stability [107], and polyadenylation [108, 109]. It also stimulates translation initiation at picornavirus internal ribosome entry site (IRES) elements [110, 111]. HnRNP L contains four RRMs that specifically recognize CA-repeat and CA-rich RNA elements [112]. It participates in intronless mRNA export [113, 114], translational regulation [98], mRNA stability [112, 115], poly(A) site selection, and alternative splicing [112]. HnRNP L competes with microRNAs for binding to a CA-rich RNA element within the *vegfa* (vascular endothelial growth factor A) 3′ UTR [116]. Recently, activities of hnRNP L were analyzed on a genome-wide level, and an *in vivo* enrichment of CA motifs as hnRNP L binding sites was confirmed. A position-dependent splicing regulation was demonstrated: while binding to intronic regions upstream of alternative exons leads to repressed splicing, binding to the downstream intron activates splicing [117].

Concerning their role as splicing regulators, many examples of hnRNPs and hnRNP-like proteins show negative regulation, including Nova 1 [118], hnRNP A1 [119], Fox2 [120], HuR [121], hnRNP H [122], hnRNP F [123], and PTB [124, 125]. A positive regulation has been shown for hnRNP A1 [126], hnRNP H [127], hnRNP G [128], and PTB [129]. Similar to SR proteins, hnRNPs show a position-dependent effect on splicing regulation [130].

7.5.3 Mechanisms of Splicing Regulation

It is frequently difficult to make a clear distinction between "constitutive" and "alternative" splicing. The decision depends on *cis*-acting elements like strong or weak splice sites (a higher degree of similarity to the consensus sequence increases splice site strength) and additional enhancer or silencer elements in the vicinity of the splice sites. Furthermore, the abundance and concentration of each splicing factor in a given cell type affects the splicing decision [131–133].

SR proteins are the main enhancers known to facilitate splice site recognition and exon inclusion by binding to ESEs. In general, they help components of the spliceosome to bind the pre-mRNA. This includes the recruitment of the U1

snRNP to the 5′ splice site. Additionally, they recruit the U2AF heterodimer and U2 snRNP to the 3′ splice site and help establish exon definition complexes (see Section 7.2.4) [76, 134, 135]. Their activity is mediated by their RS domain and the phosphorylation status (reviewed in [86]). The enhancing effect SR proteins exert on exon inclusion is position dependent. Binding of SR proteins to intronic regions can induce exon skipping [136]. Furthermore, binding of SR proteins to exonic regions can have a differential impact on the inclusion of cassette exons. Binding of SR proteins to ESEs within the cassette exon enhances its inclusion, whereas binding to ESEs within the flanking constitutive exons promotes skipping of the cassette exon [137, 138].

SR proteins can cooperate to promote exon inclusion. Different SR proteins can recognize the same ESE and compensate for each other or act cooperatively by binding to adjacent ESEs [69]. Additionally, SR proteins may form larger complexes with other RS domain-containing proteins, such as the SR-related nuclear matrix proteins SRm160 and SRm300, which are unable to bind RNA by themselves. These coactivators can form multiple interactions with snRNPs and enhancer-bound SR proteins; thus they enhance activity through bridging interactions between ESEs and spliceosomal components [139].

The splicing process can be inhibited by various mechanisms. Often, hnRNPs like PTB or hnRNP A1 are involved. The simplest way of inhibition is sterically blocking positive regulators. This happens when silencer elements are located closely to splice sites or to splicing enhancer elements, so that splicing is inhibited by blocking the access of snRNPs or positive regulatory factors. PTB, for example, binds the polypyrimidine tract and therefore blocks binding of U2AF to alternatively spliced exons [125]. Several other mechanisms by which PTB inhibits splicing have been elucidated [140]. It can inhibit U2AF binding also when bound to exonic sequences [141]. PTB binding to ISSs can inhibit the transition from an exon definition to an intron definition complex [142] or prevent interaction of the U1 snRNP with other spliceosomal components [143]. Furthermore, PTB might induce exon skipping by looping out exons flanked by intronic PTB binding sites [144].

Like SR proteins, hnRNPs can cooperate to inhibit exon inclusion [68]. Recently, it was shown that inclusion of the *cd45* exon 4 is repressed by hnRNP L binding to an ESS. HnRNP L recruits hnRNP A1 and together the two hnRNPs induce extended contacts of the U1 snRNP with exonic sequences, preventing U6 snRNP contacts with the 5′ splice site and subsequent spliceosomal catalysis [145].

Splicing of individual pre-mRNAs usually involves the integration of additive and competitive signals from both splicing activating and repressing elements. Along this line, SR proteins can induce exon inclusion by competing with repressing hnRNPs. One example is the role of hnRNP A1 in the repression of exon 3 of the HIV1 *tat* pre-mRNA. An ESS in exon 3 binds the repressor hnRNP A1 with high affinity and inhibits splicing by propagating the binding of further hnRNP A1 proteins toward the 3′ splice site. This propagative binding can be inhibited by the binding of the SR protein SRSF1 to an upstream ESE, which then activates splicing. Additionally, hnRNP A1 binds an ISS located upstream of exon 3, thereby preventing binding of the U2 snRNP [7, 119, 146].

It is important to note that the same splicing factors can stimulate as well as inhibit the inclusion of cassette exons depending on their respective binding site position. This has been shown for SR proteins (see previous text) and hnRNPs. For example, hnRNP H has been shown to promote exon inclusion when bound to intronic positions, but induce exon skipping when bound to exonic sequences [147, 148]. The hnRNP-like splicing factor Nova1 is exclusively expressed in CNS neurons and recognizes YCAY clusters. A genome-wide map revealed that the position of its binding site relative to the regulated exon dictates if Nova1 promotes exon inclusion or skipping [118].

The accumulated knowledge on the impact of *cis*-regulatory motifs, exon features (e.g., length, splice site strength), and RNA structure was successfully combined to build a "splicing code" that accurately predicts tissue-specific expression of alternatively spliced cassette exons [149].

7.5.4 Transcription-Coupled Alternative Splicing

Splicing is not only controlled by a plethora of different splicing factors, but it is also coupled to transcription, already shown in early studies [150]. Global sequencing analyses of multiple tissues and cell types in different organisms indicate that co-transcriptional splicing is widespread ([25, 151–157], reviewed in [158]). In budding yeast, fly, and human cell lines and tissues, the vast majority of introns are co-transcriptionally spliced [25, 151, 154, 156, 157]. Due to their experimental and analytical differences, it is sometimes hard to compare the studies. While some findings show that intron length negatively correlates with co-transcriptional splicing frequency in mouse, human, and fly [25, 155, 156], another study, focusing on highly expressed genes with long introns, came to the exact opposite conclusion [151]. However, numerous studies agree that constitutive splicing occurs to a greater degree in a co-transcriptional manner than alternative splicing [25, 151, 155, 156]. One study in mouse macrophages found that full-length yet incompletely spliced transcripts accumulated in the chromatin fraction [152]. The relatively low frequency of co-transcriptional splicing in this and in another mouse study [155] is in contrast to the high numbers found in yeast, fly, and human cells. To provide clear evidence, analysis of directly comparable human and mouse cell types should be addressed.

The alternative splicing decision can be influenced by several elements, including promoters [159, 160], transcription factors [161, 162], and coactivators [163–165], as well as transcription enhancers [166], chromatin remodelers [167], and factors affecting chromatin structure [168–172]. Two models are currently discussed that are not mutually exclusive: the recruitment model and the kinetic model (reviewed in [173]).

The recruitment model involves the recruitment of splicing factors to transcription sites by the transcription machinery. The carboxy-terminal domain (CTD) of RNA polymerase II (Pol II) has a key role in functionally coupling transcription to capping and 3′ processing. Additionally, several alternative splicing factors associate with the CTD, implicating this domain in alternative splicing. One example is the splicing factor SRSF3, which interacts with the CTD and inhibits inclusion of cassette exon 33 in the fibronectin mRNA [174].

The kinetic model proposes that the rate of transcription elongation affects the outcome of alternative splicing. One possibility is that upon Pol II pausing or slowing down, inclusion of alternative exons increases. An upstream exon with a weak 3′ splice site can be defined, before the downstream exon is synthesized, resulting in exon inclusion at slow transcription rates but exclusion at fast transcription kinetics. Other mechanisms include a Pol II "roadblock" upon DNA binding by proteins, like the CCCTC-binding factor (CTCF), which stalls the Pol II complex and therefore promotes inclusion of the alternative exon 5 in *cd45* [175].

As histone modifications directly affect Pol II extension speed, they can also have an impact on alternative splicing. Exons show increased nucleosome occupancy, probably caused by their higher GC content compared with the flanking intronic regions [176–178]. Furthermore, the histones associated with exons are enriched in certain modifications, which influence alternative splicing decisions (reviewed in [179]). Trimethylation of histone H3 lysine 9 (H3K9me3) is correlated with transcriptional repression. Enrichment of H3K9me3 marks on alternative exons in the *cd44* gene has been shown to increase exon inclusion [171]. The H3K9me3 modification is recognized by the chromodomain protein HP1γ, which reduces the local elongation rate of Pol II. Conversely, an increase in the Pol II transcription rate by increased histone 3 lysine 9 acetylation (H3K9ac) leads to skipping of the *ncam* exon 18 [172].

Similar to the recruitment model discussed earlier, proteins recognizing specific histone modifications have been shown to modulate alternative splicing by recruitment of splicing factors. One example is trimethylation of histone 3 lysine 36 (H3K36me3). This mark can be recognized by the Mrg15 (MORF-related gene 15) protein, which recruits PTB to an ISS near a mutually exclusive exon in *fgfr2* (fibroblast growth factor 2), repressing its inclusion in mesenchymal cells [169]. Furthermore, it has been proposed that the short isoform of Psip1 (PC4 and SF2 interacting protein 1) enhances exon inclusion by recruitment of the splicing factor SRSF1 to H3K36me3 marks [170].

7.5.5 Alternative Splicing and Nonsense-Mediated Decay

Apart from increasing protein diversity, alternative splicing can also result in mRNA degradation via the nonsense-mediated mRNA decay (NMD) pathway.

NMD is one of several RNA surveillance mechanisms to ensure the accuracy of gene expression by degrading mRNAs that contain a premature termination codon (PTC). At first, it was thought that NMD only removes defective mRNAs arising from errors in gene expression to avoid accumulation of truncated, nonfunctional proteins [180]. Nowadays, it is known that alternative splicing can introduce PTCs and exploit NMD to achieve quantitative posttranscriptional regulation [181]. In mammals, a stop codon is recognized as premature if it is located >50–55 nucleotides upstream of an exon–exon junction, which is marked by an accumulation of several proteins and called an exon junction complex (EJC) [182]. According to this rule, one third of the human alternative mRNA isoforms in the RefSeq database were predicted to be subject to NMD [183]. However, upon siRNA-mediated depletion of the NMD factor UPF1 in HeLa

cells, only 10% of PTC-containing genes were slightly increased in their mRNA levels [184, 185]. Such a rather minor role is in accordance with a microarray study, showing uniformly low levels of PTC-containing splice variants across diverse mammalian cell types and tissues [186].

Although the usage of alternative splicing coupled to nonsense-mediated mRNA decay (AS-NMD) might be less prevalent than indicated by initial computational surveys, this process is pivotal in regulating the expression of certain gene families. Among other RNA-binding proteins, AS-NMD was shown to be prevalent for members of the SR protein and hnRNP families, indicating that it is an important mechanism for the homeostatic regulation of splicing factors [80, 187, 188].

Further, recent work has shown the function of NMD during cellular differentiation and in response to stress, regulating the expression of certain splicing isoforms (reviewed in [189]). Coupled to the observation that the deletion of NMD factors is embryonic lethal in mouse [190–193], these findings emphasize the importance of this mRNA surveillance mechanism for the maintenance of physiological processes.

7.5.6 Alternative Splicing and Disease

Aberrant splicing has been recognized as the cause of several diseases and also appears to drive cancer progression [194–196]. 15% of the known disease-causing single nucleotide polymorphisms (SNPs) are located within splice sites, and >20% in predicted splicing elements [197, 198]. A comprehensive list of diseases caused by mutated 5′ and 3′ splice sites including cystic fibrosis, Alzheimer's disease, and several types of cancer is available at the database for aberrant splice sites (DBASS) [199].

Mutations of *cis*-acting elements can result in several aberrant splicing events: mutations disrupting exon definition, for example, in ESEs, 5′ or 3′ splice sites, often lead to exon skipping, resulting in nonfunctional proteins, or in the case of frame shifting to the introduction of PTCs. One example for disrupted exon definition is spinal muscular atrophy, which is described later. Similar effects are seen with mutations that activate cryptic splice sites, resulting in the retention of intronic sequences. Such activation of cryptic splice sites was already described in 1982 to cause β-thalassemia [200]. Furthermore, mutations in silencer or enhancer elements affecting the inclusion ratio of cassette exons do not alter the encoded mRNA/protein isoforms, but nevertheless can induce pathological effects as isoform ratios are important cell-type-specific determinants. For example, several intronic SNPs in the neuregulin receptor *erbB4* are associated with the increased expression of splicing isoforms upregulated in patients with schizophrenia [201].

Mutations in *trans*-acting factors can also have a severe impact on splicing regulation. Consistently, their occurrence in core spliceosomal factors is very rare, suggesting that mutations with an impact on the basal splicing machinery are embryonic lethal. The few known examples include mutations in the splicing factor SF3B1 (a component of the U2 snRNP) that are frequently observed in leukemia patients [202]. In contrast, mutations in splicing factors important for

alternative splicing are more frequent. Examples are mutations in TDP-43 and FUS, which are connected to amyotrophic lateral sclerosis (ALS) and other neurodegenerative disorders [203, 204].

Consequently, modulation of disease-causing aberrant splicing is used as a therapeutic approach [196, 205]. In spinal muscular atrophy, a motor neuron disease, the *smn2* gene is the only source for the essential survival motor neuron (SMN) protein due to an inactivation of *smn1*. Inefficient inclusion of exon 7 in the *smn2* mRNA, due to a silent mutation disrupting an ESE, leads only to the production of residual amounts of full-length protein [206]. Antisense oligonucleotides (ASOs) have been developed that force inclusion of exon 7 by masking a downstream ISS. ASOs are small oligonucleotides that base pair with exons, splice sites, or splicing factor binding sites to subsequently modulate splicing decisions [205]. This leads to the increased production of functional SMN protein, resulting in enhanced motor neuron function and survival (from 10 to >500 days) in a mouse model of severe disease [207]. One of these ASOs is now the first antisense drug functioning via splicing correction and the first FDA-approved treatment for SMA [208]. In general, ASOs show high efficacy, delivery to several tissues, the ability to cross the blood–brain barrier and, until now, no severe side effects, making them promising new therapeutics for the treatment of splicing-related diseases.

7.6 Controlled Splicing in *S. cerevisiae*

7.6.1 Alternative Splicing

Alternative splicing events in *S. cerevisiae* are rare with only three examples known so far. The most extensively alternatively spliced gene is the nuclear export factor *mtr2* [33]. *Mtr2* contains an intron in its 5′ UTR, which includes two 5′ splice sites and three 3′ splice sites. Five of the six possible combinations and the unspliced transcript are detectable. The six different transcripts either encode proteins with different N-termini or 5′ UTRs containing differing numbers (up to three) of upstream open reading frames (uORFs). The function of these different encoded proteins/5′ UTRs or how splice site selection is regulated is unknown.

A further example for alternative splice site usage in *S. cerevisiae* is *src1*. SRC1 acts in sub-telomeric gene expression and TREX-dependent mRNA export [209]. Its intron contains two overlapping 5′ splice sites: GCAA**GUGAGU** (No. 1 underlined, No. 2 bold [210]). Usage of the downstream 5′ splice site results in the expression of a long protein isoform that codes for two transmembrane domains [209]. Usage of the upstream 5′ splice site results in a shorter protein with only one transmembrane domain and reduced activity. Again, it is not known how (and if) splice site selection is regulated.

In *S. cerevisiae*, three SR-like homologs (NPL3, HRB1, and GBP2) and one hnRNP-like protein (HRP1) have been identified. Mutagenesis studies indicate that only NPL3 may be involved in splicing [211]. However, RNA-binding proteins important for the splicing of individual transcripts have been reported.

Mer1 is transcribed only during meiosis and activates splicing by binding to an intronic enhancer sequence (AYACCCUY) near the 5′ splice site. Splicing activation by MER1 is further dependent on a reduced basal splicing efficiency of the introns (e.g., a non-consensus 5′ splice site) and on the NAM8 protein, which is part of the U1 snRNP [212].

Further on, intron retention can be regulated by a negative feedback loop of the encoded protein: overexpression of the RNA export factor YRA1 is toxic to cells. Therefore, its expression has to be tightly regulated. YRA1 restricts its own expression by inhibiting the splicing of a highly unusual intron in its ORF. With 766 nucleotides, this intron is very large: it is located 300 nucleotides downstream of the AUG and contains a non-consensus branch point (GACUAAC). All these unusual features seem to be important for autoregulation, which relies on a suboptimal splicing efficiency and co-transcriptional binding of YRA1 [213, 214]. The unspliced pre-mRNA is exported to the cytoplasm, where its degradation is initiated by EDC3-activated decapping and completed by XRN1 digestion.

In addition to these cases, in which a specific protein regulates one (or four) specific transcripts, the spliceosome itself might differentiate between different introns. Genome-wide studies of changes in splicing efficiency after the introduction of mutations in 18 core spliceosomal components revealed several transcript specific effects [215]. This implies that not only specialized factors but also the core spliceosome machinery itself can influence differential splicing decisions.

7.6.2 Regulated Splicing

Instead of alternative splicing, "regulated splicing" is predominantly found in *S. cerevisiae*. There, nonfunctional introns are retained in the mature mRNA, introducing PTCs that ultimately lead to mRNA decay. The degradation of unspliced pre-mRNAs can occur in the nucleus involving the exosome. Additionally, intron-containing mRNAs can be exported to the cytoplasm, where they are degraded by either the 5′ to 3′ exonuclease XRN1 or the NMD pathway. The decision, if an intron-containing mRNA is directed to the NMD pathway, depends on the intron's identity [216].

The most prominent example of regulated splicing in *S. cerevisiae* occurs during meiosis. All 13 of the intron-containing genes related to meiosis are spliced inefficiently during exponential growth in rich medium, but splicing is dramatically induced during sporulation [217]. This regulation mechanism seems to depend on the competition of meiosis-related genes with intron-containing ribosomal proteins for the splicing machinery [218]. During meiosis, the expression of ribosomal proteins is temporary repressed. During this time period, the global splicing efficiency, including splicing of meiosis-related genes, is improved. Ribosomal proteins comprise ~90% of all splicing substrates during vegetative growth, outcompeting other intron-containing pre-mRNAs for the splicing apparatus. Therefore transcriptional repression of ribosomal proteins leads to an overall change in the composition of nuclear pre-mRNAs, ultimately allowing for efficient splicing of otherwise inefficiently spliced meiosis-related pre-mRNAs.

7.6.3 Function of Splicing in *S. cerevisiae*

In contrast to the finding that introns seem to be beneficial for high gene expression, most of them can be deleted without affecting growth in rich medium [219, 220]. Also, multiple deletions of introns in one strain, for example, all 16 introns within the 15 intron-containing cytoskeleton-related genes, showed no impact on growth under standard laboratory conditions. Only the deletion of introns in RNA-binding proteins caused growth defects in rich medium. There, introns seem to be important for gene expression by the endogenous promoter, as heterologous expression of the genes from an *act1* promoter restored cell growth.

Introns also seem to be important for fine-tuning gene expression and growth under stress conditions. Parenteau *et al.* systematically deleted the introns of all ribosomal proteins and investigated their impact on gene expression, rRNA maturation, and growth under stress conditions [219]. They found that 21% of the intron deletions inhibited growth in the presence of staurosporine and 37% affected growth during at least one of five stress conditions tested. Furthermore, intron deletions did not only alter the expression of the respective gene but also affected pre-rRNA processing and expression of the paralog in duplicated genes [219]. This shows that, albeit in most cases, the few remaining introns in *S. cerevisiae* are not important for cell survival *per se*, but they do play a role in posttranscriptional gene regulation and therefore might not be readily expelled from the genome in future evolution.

Recently, a novel role for the spliceosome in the regulation of intronless genes has been discovered [221]. Intronless genes containing consensus 5′ splice sites and branch point sequences are bound by the spliceosome and spliced (at least the first step of splicing is performed). The incorrectly spliced pre-mRNA is subsequently degraded, leading to the downregulation of gene expression. The authors suggest that the expression of ~1% of the intronless genes in *S. cerevisiae* is regulated by this so-called spliceosome-mediated decay (SMD) mechanism.

The advancement of novel high-throughput sequencing methods allows for an unprecedented in-depth analysis of expressed isoform variants. Consequently, several recent transcriptomic studies discovered novel introns and usage of alternative splice sites in *S. cerevisiae*, significantly expanding the role of splicing in this "intron-poor" eukaryote [222–224].

7.7 Splicing Regulation by Riboswitches

A decade ago, a novel RNA-based regulatory mechanism was discovered. The so-called riboswitches are structured RNA elements usually residing in the 5′ UTR of bacterial genes [225]. Riboswitches consist of two domains: an aptamer domain and an expression platform. These aptamer domain senses the amount of a small molecule ligand. Its binding leads to a structural rearrangement, which is translated to the expression platform, subsequently modulating gene expression. Most bacterial riboswitches regulate gene expression by either transcriptional termination or translational repression. They are found predominantly in genes related to the metabolic pathways of their cognate ligand.

During the last years, a plethora of different riboswitch classes sensing a diverse set of ligands has been discovered. In most cases the ligands are nucleobases, amino acids, or coenzymes, but also ion- and second messenger-sensing riboswitches have been identified (reviewed in [225–227]).

7.7.1 Regulation of Group I Intron Splicing in Bacteria

Recently two structurally different classes of riboswitches sensing the second messenger cyclic diguanylate (c-di-GMP) were discovered [228, 229]. In *Clostridium difficile*, a class II c-di-GMP riboswitch was identified upstream of a group I self-splicing intron [228]. Here, the start codon of the downstream gene is engaged in base pairing interactions with sequences of the group I intron structure. Furthermore, the intron contains an atypical 5′ splice site that is partially sequestered by base pairing with an anti-5′ splice site sequence. In the absence of c-di-GMP, guanosine triphosphate (GTP) cannot attack the sequestered 5′ splice site, but attacks a site near the 3′ splice site, resulting in a nonfunctional mRNA, with an accessible start codon, but lacking a ribosomal binding site (rbs) (see Figure 7.3a, middle). Formation of the riboswitch structure in the presence of c-di-GMP leads to disruption of the anti-5′ splice site stem. The correct 5′ splice site can now be attacked by GTP and the group I intron removed completely. As a result, the start codon is not longer sequestered. In addition, a functional rbs is created by the joining of the exon sequences (see Figure 7.3a, left side). Both events ultimately lead to gene expression.

Apart from the allosteric activation of self-splicing, the c-di-GMP riboswitch in *C. difficile* can regulate translation of the downstream gene in a second step [230]. After removal of the group I intron, the upstream riboswitch lies next to the newly created rbs. Access to the rbs is then regulated by binding of c-di-GMP to the aptamer domain like in classical translation-controlling riboswitches (see Figure 7.3, right side and bottom). To this date, this is the only example of a riboswitch regulating a self-splicing intron.

7.7.2 Regulation of Alternative Splicing by Riboswitches in Eukaryotes

So far only one riboswitch class – the thiamine pyrophosphate (TPP) riboswitch – has been found in all three domains of life. In all identified cases, the eukaryotic riboswitches are located in introns and regulate alternative splicing in a TPP-dependent manner. Depending on the organism, the intronic sequences containing the riboswitch are located in different parts of the pre-mRNA, and alternative splicing subsequently triggers different downstream effects [231, 232]. In *Aspergillus oryzae*, a TPP riboswitch resides in an intron in the 5′ UTR of a thiamin biosynthetic gene [233]. In *Neurospora crassa*, three TPP riboswitches have been identified [234]. Two of them reside in an intron in the 5′ UTR (see Figure 7.3b). There, the intron encodes two 5′ splice sites. Under conditions of low TPP concentration, the upstream 5′ splice site is used, leading to the complete removal of the intronic sequence and high levels of gene expression. When the TPP concentration is high, the downstream 5′ splice site mediates partial retention of the intronic sequence. The retained sequence introduces a uORF,

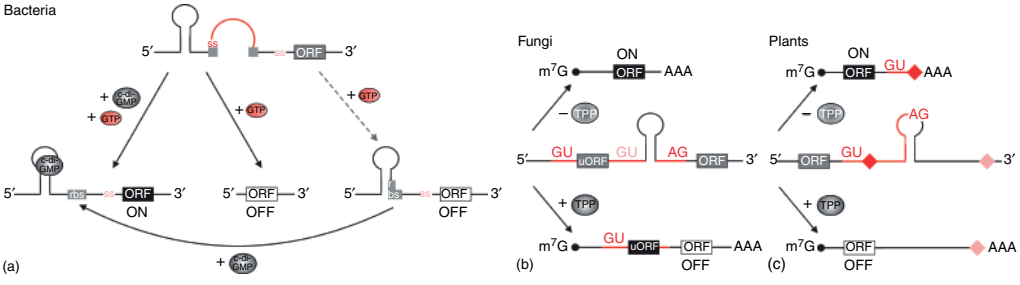


Figure 7.3 Splicing regulation by riboswitches. (a) In the bacterium *C. difficile*, a c-di-GMP binding riboswitch regulates splicing of a group I intron. Depending on the presence of the ligand, different mRNAs are produced by alternative splicing. Left: Upon c-di-GMP binding to the riboswitch, an otherwise sequestered 5′ splice site (indicated in red) becomes accessible to the cofactor GTP, and the complete group I intron is removed. Therefore, joining of the exon sequences creates an accessible ribosomal binding site (rbs) and the downstream gene can be expressed. Middle: In the absence of c-di-GMP, the correct 5′ splice site (indicated in red) is inaccessible for GTP attack. GTP attack on a downstream site (indicated in pink) occurs, creating a truncated mRNA without a ribosomal binding site. Therefore, gene expression is inhibited. Right: In very rare cases, the group I intron is correctly spliced in the absence of c-di-GMP. Nevertheless, gene expression does not occur in the absence of c-di-GMP, as the newly created rbs is sequestered within the riboswitch. In cases where the complete group I intron has been removed, subsequent c-di-GMP binding to the aptamer domain of the riboswitch leads to structural rearrangements, rendering the rbs accessible. Gene expression can thus be switched on or off, depending on the ligand binding state of the riboswitch. (b) In the filamentous fungus *N. crassa*, two genes harbor a TPP riboswitch within an intron in their 5′ UTR. Both introns contain two 5′ splice sites. Top: In the absence of TPP, the downstream 5′ splice site (pink) is sequestered by base pairing interactions with the free TPP aptamer domain. Consequently, the upstream 5′ splice site (red) is used, leading to complete intron removal and subsequently to gene expression. Bottom: In the presence of TPP, the aptamer domain binds its ligand, which renders the downstream 5′ splice site accessible to the spliceosome. Thus, a part of the intron is retained after splicing, introducing a uORF, which inhibits gene expression. (c) In higher plants (e.g., *Arabidopsis thaliana*), TPP aptamer domains are found in introns within 3′ UTRs. There, gene expression is regulated by usage of two different 3′ processing sites (diamonds). Top: In the absence of TPP, the 5′ splice site is sequestered by the aptamer domain, leading to intron retention. 3′ processing occurs at the upstream site (red) encoded within the intron. This leads to a stable mRNA with a short 3′ UTR and gene expression. Bottom: In the presence of TPP, the 5′ splice site is accessible and the upstream 3′ processing site is removed along with the intron. Usage of the downstream 3′ processing site (pink) leads to an mRNA with a long 3′ UTR. This mRNA is unstable, as long 3′ UTRs in plants trigger NMD. Therefore, gene expression is repressed. AAA = poly(A) tail, c-di-GMP = cyclic diguanylate, GTP = guanosine triphosphate, m7G = 7-methylguanosine cap, ORF = open reading frame, rbs = ribosomal binding site, ss = splice site, TPP = thiamine pyrophosphate, uORF = upstream open reading frame.

which negatively affects gene expression. In the third case, the intron containing the riboswitch is located in the main ORF of the gene. When the TPP level is low, the intron is removed completely, resulting in gene expression [235]. TPP binding leads to incomplete intron removal by usage of downstream 5′ splice sites, disrupting the main ORF by the introduction of frame shifts. So, in all three cases, TPP leads to the downregulation of gene expression by modulation of 5′ splice site usage.

In higher plants, TPP riboswitches reside in the 3′ UTRs and control intron retention by regulating the accessibility of the 5′ splice site (see Figure 7.3c) [236, 237]. At low TPP concentrations, the 5′ splice site is inaccessible and a stable mRNA with a short 3′ UTR is expressed. In the presence of high amounts of TPP, the 5′ splice site is accessible and the intron in the 3′ UTR is removed. Splicing of the intron also removes the major 3′ end processing site. As a result, another downstream 3′ processing site is used, leading to an elongated 3′ UTR that induces degradation by NMD.

In all cases studied, sequences within the aptamer domain of the TPP riboswitch base pair with splicing signals (usually the 5′ splice site), rendering them inaccessible for the spliceosome. This sequestration of splicing signals then triggers the usage of alternative 5′ splice sites, exon skipping, or intron retention. As the base pairing sequences in the TPP riboswitch are part of the ligand binding pocket, binding of TPP leads to structural rearrangements, which renders the splicing signals accessible. Subsequent downstream events then repress gene expression. This is achieved either by translational repression due to uORFs in the 5′ UTR or by triggering NMD via PTCs or the length of the 3′ UTR.

The TPP riboswitch in *N. crassa* located in the intron of the main ORF is an interesting exception (see preceding text). Here, the base pairing interactions in the absence of TPP do not regulate alternative splicing by 5′ splice site sequestration, but facilitate intron removal [235]. This is achieved by a long-range interaction between the aptamer domain and several conserved nucleotides downstream of the 5′ splice site. It seems that, upon structure formation, reducing the effective distance between the 5′ and 3′ splice sites enhances splicing efficiency.

Until now, no eukaryotic homologs have been identified for the other riboswitch classes discovered in bacteria and archaea. Still, the report of a putative arginine binding riboswitch, present in an intron in the 5′ UTR of an arginase in the fungus *Aspergillus nidulans*, suggests that other eukaryotic riboswitches might exist [238].

7.8 Splicing and Synthetic Biology

7.8.1 Impact of Introns on Gene Expression

Splicing is tightly linked with all stages of mRNA metabolism, including transcription, mRNA processing, nuclear export, and translation. Intron sequences may harbor transcriptional regulatory elements or affect DNA accessibility by determining nucleosome arrangement, influence export processes, mRNA

stability, and the translatability of the mRNA (reviewed in [239]). As a consequence, the expression of a gene can change dramatically depending on the presence or absence of introns.

Gene expression is usually reduced upon removal of endogenous introns [240]. So the maintenance of introns can lead to considerable higher protein expression, for example, for overexpression studies [241]. It is exemplified in Figure 7.4a, where the expression of the MAX (Myc associated factor X) protein from an intronless cDNA was compared with a cDNA retaining one endogenous intron.

Also the heterologous expression of transgenes can be increased significantly by adding just a single generic intron [242–245]. The extent of the effect depends on intron identity, intron position within the gene, and the surrounding exonic sequences. Placing the same intron between different exons may yield opposing results [241, 246]. The insertion of intron 2 of the β-globin gene into a firefly luciferase reporter gene increased its expression 3-fold, the insertion of a synthetic intron only 1.5-fold (see Figure 7.4b). In contrast, insertion of β-globin intron 1 led to undetectable reporter activity. Therefore, the inability of β-globin intron 1 to confer efficient splicing in a heterologous context is apparently due to its weak splicing signals and missing enhancing sequences in the artificial context [247]. The insertion of two short introns from the immunoglobulin heavy chain into both a green fluorescent protein (GFP) reporter gene and a Cre recombinase cDNA increased gene expression up to 30-fold in CHO cells. These introns were chosen because they were short, compatible with high levels of gene expression, and without evidence of containing regulatory sequences [248]. In line with this approach, several commercially available expression vectors also contain short synthetic introns in their 5′ UTRs known to enhance the stability of the mRNA by influencing polyadenylation [249].

7.8.2 Control of Splicing by Engineered RNA-Based Devices

The controlled removal of intronic sequences offers the possibility to engineer user-defined gene expression systems. RNA-based control devices generally couple *in vitro* selected RNA aptamers as sensory domains to functional RNA domains (like a rbs, splice site, or a ribozyme). By modulating the accessibility of elements essential for splicing, such as the 5′ splice site, the branch point, or the 3′ splice site, engineered riboswitches have been shown to control both constitutive and alternative splicing.

In a pioneering study, a theophylline-binding aptamer was inserted close to a 3′ splice site. The addition of the ligand theophylline resulted in a 4-fold reduction of gene expression in an *in vitro* splicing assay [250, 251]. The data indicated that theophylline binding specifically blocked the recognition of the 3′ splice site. This aptamer was also used to modulate splicing efficiency by including the branch point sequence into the aptamer sequence [252]. In the presence of theophylline, the downstream exon was skipped twice more often than in its absence, indicating that engineered riboswitches can also modulate and therefore investigate the impact of alternative splicing.

A tetracycline-binding aptamer was used to regulate pre-mRNA splicing in yeast [253]. The aptamer was inserted into a yeast intron in close proximity to

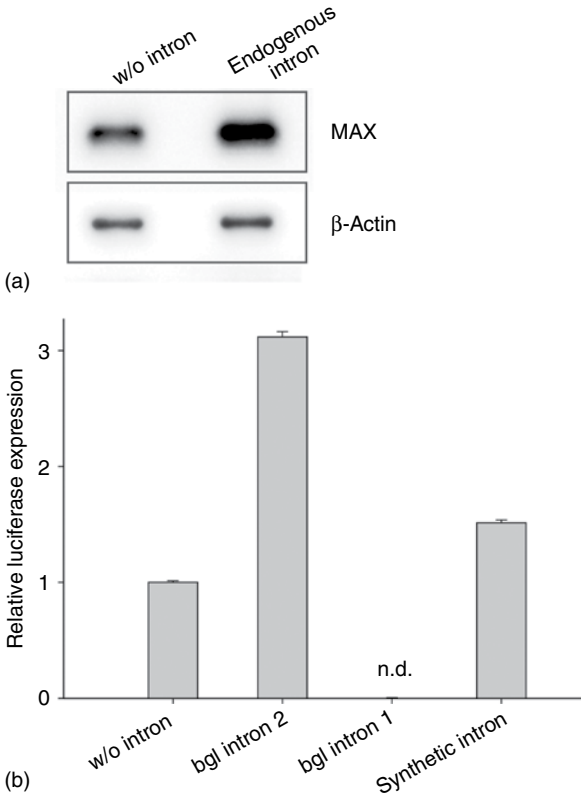


Figure 7.4 Influence of introns on gene expression. (a) Western blot analyses after overexpression of MAX protein either lacking (w/o) or containing one endogenous intron in the ORF. Overexpression was performed in HeLa cells and protein was isolated after 24 h. β-Actin was used as a loading control. (b) Firefly luciferase reporter gene constructs without (w/o) intron or containing the β-globin introns 1, 2 or a synthetic intron. Firefly luciferase activity was measured in triplicates 24 h after transfection of HeLa cells using *Renilla* luciferase as transfection control. bgl = β-globin, n.d. = not determined.

either the 5′ splice site or the branch point. Maximal regulation (16-fold) was observed with a construct in which the 5′ splice site was masked by intramolecular base pairing when placed within the closing stem of the aptamer. This blocked the accessibility of the 5′ splice site to the U1 snRNP. The dynamic range of regulation was increased by additionally inserting a second aptamer-containing intron.

Another programmable control device expands the possibility to engineer alternative splicing by being triggered by the presence of specific protein binding to an aptamer located in the intronic sequence. This approach has been successfully used to rewire both the Wnt and nuclear factor κB signaling pathways in mammalian cells [254]. In plants a naturally occurring rRNA-mimicking structure was used to regulate cassette exon splicing in response to the expression of a ribosomal protein. By using an engineered variant of the RNA structure from another plant species highly efficient, orthogonal gene activation could be achieved in *Nicotiana benthamiana* [255].

Up to now, the number of artificially engineered systems used to control pre-mRNA splicing is limited, but the examples presented impressively demonstrate that synthetic devices have an immense potential for controlling splicing and, thus, both level and identity of target gene expression.

7.9 Conclusion

Besides the importance of splicing for increasing proteome diversity, there is a clear impact of introns on gene expression levels with introns often stimulating but sometimes also reducing gene expression. There is no universal requirement for introns, but their presence has to be carefully considered during the *de novo* design of genetic pathways. Moreover, given the great importance of RNA splicing for gene regulation *per se*, RNA elements that target splicing may soon provide general and highly applicable platforms for engineering gene regulation systems.

Acknowledgments

This work has been supported by grants of the Deutsche Forschungsgemeinschaft (CRC902 A2 B.S. and B14 J.E.W.) and the Else Kröner-Fresenius-Stiftung (J.E.W.).

Definitions

Splicing removal of intronic sequences from the pre-mRNA
Exon sequences of a gene included into the mature mRNA
Intron intervening sequences removed upon splicing
Spliceosome machinery that removes introns from the pre-mRNA
Alternative splicing generation of multiple mature mRNA molecules from a single gene

References

1 Berget, S.M., Moore, C., and Sharp, P.A. (1977) Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 3171–3175.
2 Chow, L.T., Roberts, J.M., Lewis, J.B., and Broker, T.R. (1977) A map of cytoplasmic RNA transcripts from lytic adenovirus type 2, determined by electron microscopy of RNA:DNA hybrids. *Cell*, **11**, 819–836.
3 Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
4 Burge, C.B., Tuschl, T., and Sharp, P.A. (1999) 20 Splicing of precursors to mRNAs by the spliceosomes, in *The RNA World*, CSHL Press, pp. 525–560.
5 Will, C.L. and Luhrmann, R. (2011) Spliceosome structure and function. *Cold Spring Harbor Perspect. Biol.*, **3**, a003707.

6 Sheth, N., Roca, X., Hastings, M.L., Roeder, T. *et al.* (2006) Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res.*, **34**, 3955–3967.

7 Black, D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, **72**, 291–336.

8 Moore, M.J., Query, C.C., and Sharp, P.A. (1993) 13 Splicing of precursors to mRNA by the spliceosome, in *The RNA World*, CSHL Press, pp. 303–357.

9 Chen, M. and Manley, J.L. (2009) Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat. Rev. Mol. Cell Biol.*, **10**, 741–754.

10 Will, C.L. and Lührmann, R. (2006) 13 Spliceosome structure and function,.

11 Wahl, M.C., Will, C.L., and Luhrmann, R. (2009) The spliceosome: design principles of a dynamic RNP machine. *Cell*, **136**, 701–718.

12 Brow, D.A. (2002) Allosteric cascade of spliceosome activation. *Annu. Rev. Genet.*, **36**, 333–360.

13 Matlin, A.J., Clark, F., and Smith, C.W. (2005) Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.*, **6**, 386–398.

14 Staley, J.P. and Woolford, J.L. Jr. (2009) Assembly of ribosomes and spliceosomes: complex ribonucleoprotein machines. *Curr. Opin. Cell Biol.*, **21**, 109–118.

15 Bertram, K., Agafonov, D.E., Liu, W.T., Dybkov, O. *et al.* (2017) Cryo-EM structure of a human spliceosome activated for step 2 of splicing. *Nature*, **542**, 318–323.

16 Wan, R., Yan, C., Bai, R., Huang, G. *et al.* (2016) Structure of a yeast catalytic step I spliceosome at 3.4 A resolution. *Science*, **353**, 895–904.

17 Yan, C., Wan, R., Bai, R., Huang, G. *et al.* (2016) Structure of a yeast activated spliceosome at 3.5 A resolution. *Science*, **353**, 904–911.

18 Fox-Walsh, K.L., Dou, Y., Lam, B.J., Hung, S.P. *et al.* (2005) The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 16176–16181.

19 Berget, S.M. (1995) Exon recognition in vertebrate splicing. *J. Biol. Chem.*, **270**, 2411–2414.

20 Hoffman, B.E. and Grabowski, P.J. (1992) U1 snRNP targets an essential splicing factor, U2AF65, to the 3′ splice site by a network of interactions spanning the exon. *Genes Dev.*, **6**, 2554–2568.

21 Reed, R. (2000) Mechanisms of fidelity in pre-mRNA splicing. *Curr. Opin. Cell Biol.*, **12**, 340–345.

22 Ke, S. and Chasin, L.A. (2010) Intronic motif pairs cooperate across exons to promote pre-mRNA splicing. *Genome Biol.*, **11**, R84.

23 Behzadnia, N., Hartmuth, K., Will, C.L., and Luhrmann, R. (2006) Functional spliceosomal A complexes can be assembled in vitro in the absence of a penta-snRNP. *RNA*, **12**, 1738–1746.

24 Schneider, M., Will, C.L., Anokhina, M., Tazi, J. *et al.* (2010) Exon definition complexes contain the tri-snRNP and can be directly converted into B-like precatalytic splicing complexes. *Mol. cell*, **38**, 223–235.

25 Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A. *et al.* (2012) Deep sequencing of subcellular RNA fractions shows splicing to be predominantly

co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.*, **22**, 1616–1625.
26 Vargas, D.Y., Shah, K., Batish, M., Levandoski, M. *et al.* (2011) Single-molecule imaging of transcriptionally coupled and uncoupled splicing. *Cell*, **147**, 1054–1065.
27 Wood, V., Gwilliam, R., Rajandream, M.A., Lyne, M. *et al.* (2002) The genome sequence of Schizosaccharomyces pombe. *Nature*, **415**, 871–880.
28 Fair, B.J. and Pleiss, J.A. (2017) The power of fission: yeast as a tool for understanding complex splicing. *Curr. Genet.*, **63**, 375–380.
29 Kaufer, N.F. and Potashkin, J. (2000) Analysis of the splicing machinery in fission yeast: a comparison with budding yeast and mammals. *Nucleic Acids Res.*, **28**, 3003–3010.
30 Neuveglise, C., Marck, C., and Gaillardin, C. (2011) The intronome of budding yeasts. *C.R. Biol.*, **334**, 662–670.
31 Jeffares, D.C., Mourier, T., and Penny, D. (2006) The biology of intron gain and loss. *Trends Genet.*, **22**, 16–22.
32 Bon, E., Casaregola, S., Blandin, G., Llorente, B. *et al.* (2003) Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. *Nucleic Acids Res.*, **31**, 1121–1135.
33 Davis, C.A., Grate, L., Spingola, M., and Ares, M. Jr. (2000) Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast. *Nucleic Acids Res.*, **28**, 1700–1706.
34 Lopez, P.J. and Seraphin, B. (1999) Genomic-scale quantitative analysis of yeast pre-mRNA splicing: implications for splice-site recognition. *RNA*, **5**, 1135–1137.
35 Ma, P. and Xia, X. (2011) Factors affecting splicing strength of yeast genes. *Comp. Funct. Genomics*, **2011**, 212146.
36 Ares, M. Jr., Grate, L., and Pauling, M.H. (1999) A handful of intron-containing genes produces the lion's share of yeast mRNA. *RNA*, **5**, 1138–1139.
37 Juneau, K., Miranda, M., Hillenmeyer, M.E., Nislow, C. *et al.* (2006) Introns regulate RNA and protein abundance in yeast. *Genetics*, **174**, 511–518.
38 Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W. *et al.* (2003) Global analysis of protein expression in yeast. *Nature*, **425**, 737–741.
39 Fedor, M.J. and Williamson, J.R. (2005) The catalytic diversity of RNAs. *Nat. Rev. Mol. Cell Biol.*, **6**, 399–412.
40 Cech, T.R. (1990. Nobel lecture.) Self-splicing and enzymatic activity of an intervening sequence RNA from Tetrahymena. *Biosci. Rep.*, **10**, 239–261.
41 Hausner, G., Hafez, M., and Edgell, D.R. (2014) Bacterial group I introns: mobile RNA catalysts. *Mob. DNA*, **5**, 8.
42 McNeil, B.A., Semper, C., and Zimmerly, S. (2016) Group II introns: versatile ribozymes and retroelements. *Wiley Interdiscip. Rev. RNA*, **7**, 341–355.
43 Toor, N., Keating, K.S., Taylor, S.D., and Pyle, A.M. (2008) Crystal structure of a self-spliced group II intron. *Science*, **320**, 77–82.
44 Fica, S.M., Tuttle, N., Novak, T., Li, N.S. *et al.* (2013) RNA catalyses nuclear pre-mRNA splicing. *Nature*, **503**, 229–234.
45 Jacquier, A. (1990) Self-splicing group II and nuclear pre-mRNA introns: how similar are they? *Trends Biochem. Sci*, **15**, 351–354.
46 Rogozin, I.B., Carmel, L., Csuros, M., and Koonin, E.V. (2012) Origin and evolution of spliceosomal introns. *Biol. Direct*, **7**, 11.

47 Abelson, J., Trotta, C.R., and Li, H. (1998) tRNA splicing. *J. Biol. Chem.*, **273**, 12685–12688.

48 Pan, Q., Shai, O., Lee, L.J., Frey, B.J. *et al.* (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.

49 Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I. *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.

50 Roy, B., Haupt, L.M., and Griffiths, L.R. (2013) Review: alternative splicing (AS) of genes as an approach for generating protein complexity. *Curr. Genomics*, **14**, 182–194.

51 McManus, C.J. and Graveley, B.R. (2011) RNA structure and the mechanisms of alternative splicing. *Curr. Opin. Genet. Dev.*, **21**, 373–379.

52 Breitbart, R.E., Nguyen, H.T., Medford, R.M., Destree, A.T. *et al.* (1985) Intricate combinatorial patterns of exon splicing generate multiple regulated troponin T isoforms from a single gene. *Cell*, **41**, 67–82.

53 Hastings, K.E., Bucher, E.A., and Emerson, C.P. Jr. (1985) Generation of troponin T isoforms by alternative RNA splicing in avian skeletal muscle. Conserved and divergent features in birds and mammals. *J. Biol. Chem.*, **260**, 13699–13703.

54 Medford, R.M., Nguyen, H.T., Destree, A.T., Summers, E. *et al.* (1984) A novel mechanism of alternative RNA splicing for the developmentally regulated generation of troponin T isoforms from a single gene. *Cell*, **38**, 409–421.

55 Briggs, M.M. and Schachat, F. (1993) Origin of fetal troponin T: developmentally regulated splicing of a new exon in the fast troponin T gene. *Dev. Biol.*, **158**, 503–509.

56 Stefancsik, R., Randall, J.D., Mao, C., and Sarkar, S. (2003) Structure and sequence of the human fast skeletal troponin T (TNNT3) gene: insight into the evolution of the gene and the origin of the developmentally regulated isoforms. *Comp. Funct. Genomics*, **4**, 609–625.

57 Wang, J. and Jin, J.P. (1997) Primary structure and developmental acidic to basic transition of 13 alternatively spliced mouse fast skeletal muscle troponin T isoforms. *Gene*, **193**, 105–114.

58 Briggs, M.M. and Schachat, F. (1996) Physiologically regulated alternative splicing patterns of fast troponin T RNA are conserved in mammals. *Am. J. Physiol.*, **270**, C298–C305.

59 Brotto, M.A., Biesiadecki, B.J., Brotto, L.S., Nosek, T.M. *et al.* (2006) Coupled expression of troponin T and troponin I isoforms in single skeletal muscle fibers correlates with contractility. *Am. J. Physiol. Cell Physiol.*, **290**, C567–C576.

60 Gomes, A.V., Barnes, J.A., Harada, K., and Potter, J.D. (2004) Role of troponin T in disease. *Mol. Cell. Biochem.*, **263**, 115–129.

61 Ogut, O., Granzier, H., and Jin, J.P. (1999) Acidic and basic troponin T isoforms in mature fast-twitch skeletal muscle and effect on contractility. *Am. J. Physiol.*, **276**, C1162–C1170.

62 Pan, B.S. and Potter, J.D. (1992) Two genetically expressed troponin T fragments representing alpha and beta isoforms exhibit functional differences. *J. Biol. Chem.*, **267**, 23052–23056.

63 Schilder, R.J., Kimball, S.R., and Jefferson, L.S. (2012) Cell-autonomous regulation of fast troponin T pre-mRNA alternative splicing in response to mechanical stretch. *Am. J. Physiol. Cell Physiol.*, **303**, C298–C307.

64 Schilder, R.J., Kimball, S.R., Marden, J.H., and Jefferson, L.S. (2011) Body weight-dependent troponin T alternative splicing is evolutionarily conserved from insects to mammals and is partially impaired in skeletal muscle of obese rats. *J. Exp. Biol.*, **214**, 1523–1532.

65 Vihola, A., Bachinski, L.L., Sirito, M., Olufemi, S.E. *et al.* (2010) Differences in aberrant expression and splicing of sarcomeric proteins in the myotonic dystrophies DM1 and DM2. *Acta Neuropathol.*, **119**, 465–479.

66 Sancisi, V., Germinario, E., Esposito, A., Morini, E. *et al.* (2014) Altered Tnnt3 characterizes selective weakness of fast fibers in mice overexpressing FSHD region gene 1 (FRG1). *Am. J. Physiol. Regul., Integr. Comp. Physiol.*, **306**, R124–R137.

67 Merkin, J., Russell, C., Chen, P., and Burge, C.B. (2012) Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science*, **338**, 1593–1599.

68 Huelga, S.C., Vu, A.Q., Arnold, J.D., Liang, T.Y. *et al.* (2012) Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell Rep.*, **1**, 167–178.

69 Pandit, S., Zhou, Y., Shiue, L., Coutinho-Mansfield, G. *et al.* (2013) Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. *Mol. Cell*, **50**, 223–235.

70 Witten, J.T. and Ule, J. (2011) Understanding splicing regulation through RNA splicing maps. *Trends Genet.*, **27**, 89–97.

71 Shen, H. and Green, M.R. (2004) A pathway of sequential arginine-serine-rich domain-splicing signal interactions during mammalian spliceosome assembly. *Mol. Cell*, **16**, 363–373.

72 Shen, H. and Green, M.R. (2006) RS domains contact splicing signals and promote splicing by a common mechanism in yeast through humans. *Genes Dev.*, **20**, 1755–1765.

73 Tacke, R. and Manley, J.L. (1999) Determinants of SR protein specificity. *Curr. Opin. Cell Biol.*, **11**, 358–362.

74 Manley, J.L. and Krainer, A.R. (2010) A rational nomenclature for serine/arginine-rich protein splicing factors (SR proteins). *Genes Dev.*, **24**, 1073–1074.

75 Lin, S. and Fu, X.D. (2007) SR proteins and related factors in alternative splicing. *Adv. Exp. Med. Biol.*, **623**, 107–122.

76 Cho, S., Hoang, A., Sinha, R., Zhong, X.Y. *et al.* (2011) Interaction between the RNA binding domains of Ser-Arg splicing factor 1 and U1-70K snRNP protein determines early spliceosome assembly. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 8233–8238.

77 Fu, X.D. and Maniatis, T. (1992) The 35-kDa mammalian splicing factor SC35 mediates specific interactions between U1 and U2 small nuclear ribonucleoprotein particles at the 3′ splice site. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 1725–1729.

78 Kohtz, J.D., Jamison, S.F., Will, C.L., Zuo, P. *et al.* (1994) Protein–protein interactions and 5′-splice-site recognition in mammalian mRNA precursors. *Nature*, **368**, 119–124.

79 Roscigno, R.F. and Garcia-Blanco, M.A. (1995) SR proteins escort the U4/U6.U5 tri-snRNP to the spliceosome. *RNA*, **1**, 692–706.

80 Lareau, L.F., Inada, M., Green, R.E., Wengrod, J.C. *et al.* (2007) Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature*, **446**, 926–929.

81 Ni, J.Z., Grate, L., Donohue, J.P., Preston, C. *et al.* (2007) Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev.*, **21**, 708–718.

82 Loomis, R.J., Naoe, Y., Parker, J.B., Savic, V. *et al.* (2009) Chromatin binding of SRp20 and ASF/SF2 and dissociation from mitotic chromosomes is modulated by histone H3 serine 10 phosphorylation. *Mol. Cell*, **33**, 450–461.

83 Das, R., Yu, J., Zhang, Z., Gygi, M.P. *et al.* (2007) SR proteins function in coupling RNAP II transcription to pre-mRNA splicing. *Mol. Cell*, **26**, 867–881.

84 Lin, S., Coutinho-Mansfield, G., Wang, D., Pandit, S. *et al.* (2008) The splicing factor SC35 has an active role in transcriptional elongation. *Nat. Struct. Mol. Biol.*, **15**, 819–826.

85 Huang, Y. and Steitz, J.A. (2005) SRprises along a messenger's journey. *Mol. Cell*, **17**, 613–615.

86 Zhou, Z. and Fu, X.D. (2013) Regulation of splicing by SR proteins and SR protein-specific kinases. *Chromosoma*, **122**, 191–207.

87 Dreyfuss, G., Matunis, M.J., Pinol-Roma, S., and Burd, C.G. (1993) hnRNP proteins and the biogenesis of mRNA. *Annu. Rev. Biochem.*, **62**, 289–321.

88 Michelle, L., Barbier, J., and Chabot, B. (2012) hnRNP and hnRNP-like proteins in splicing control: molecular mechanisms and implication in human diseases, in *RNA Binding Proteins* (ed. Z.J. Lorkovic), Landes Biosciences, pp. 1–25.

89 Daubner, G.M., Clery, A., and Allain, F.H. (2013) RRM-RNA recognition: NMR or crystallography…and new findings. *Curr. Opin. Struct. Biol.*, **23**, 100–108.

90 Krecic, A.M. and Swanson, M.S. (1999) hnRNP complexes: composition, structure, and function. *Curr. Opin. Cell Biol.*, **11**, 363–371.

91 Burd, C.G. and Dreyfuss, G. (1994) Conserved structures and diversity of functions of RNA-binding proteins. *Science*, **265**, 615–621.

92 Kiledjian, M. and Dreyfuss, G. (1992) Primary structure and binding activity of the hnRNP U protein: binding RNA through RGG box. *EMBO J.*, **11**, 2655–2664.

93 Pinol-Roma, S. and Dreyfuss, G. (1992) Shuttling of pre-mRNA binding proteins between nucleus and cytoplasm. *Nature*, **355**, 730–732.

94 Jean-Philippe, J., Paz, S., and Caputi, M. (2013) hnRNP A1: the Swiss army knife of gene expression. *Int. J. Mol. Sci.*, **14**, 18999–19024.

95 Campillos, M., Lamas, J.R., Garcia, M.A., Bullido, M.J. *et al.* (2003) Specific interaction of heterogeneous nuclear ribonucleoprotein A1 with the –219T allelic form modulates APOE promoter activity. *Nucleic Acids Res.*, **31**, 3063–3070.

96 Lau, J.S., Baumeister, P., Kim, E., Roy, B. *et al.* (2000) Heterogeneous nuclear ribonucleoproteins as regulators of gene expression through interactions with the human thymidine kinase promoter. *J. Cell. Biochem.*, **79**, 395–406.

97 Xia, H. (2005) Regulation of gamma-fibrinogen chain expression by heterogeneous nuclear ribonucleoprotein A1. *J. Biol. Chem.*, **280**, 13171–13178.

98 Hamilton, B.J., Burns, C.M., Nichols, R.C., and Rigby, W.F. (1997) Modulation of AUUUA response element binding by heterogeneous nuclear ribonucleoprotein A1 in human T lymphocytes. The roles of cytoplasmic location, transcription, and phosphorylation. *J. Biol. Chem.*, **272**, 28732–28741.

99 Henics, T., Sanfridson, A., Hamilton, B.J., Nagy, E. *et al.* (1994) Enhanced stability of interleukin-2 mRNA in MLA 144 cells. Possible role of cytoplasmic AU-rich sequence-binding proteins. *J. Biol. Chem.*, **269**, 5377–5383.

100 Michael, W.M., Choi, M., and Dreyfuss, G. (1995) A nuclear export signal in hnRNP A1: a signal-mediated, temperature-dependent nuclear protein export pathway. *Cell*, **83**, 415–422.

101 Bonnal, S., Pileur, F., Orsini, C., Parker, F. *et al.* (2005) Heterogeneous nuclear ribonucleoprotein A1 is a novel internal ribosome entry site trans-acting factor that modulates alternative initiation of translation of the fibroblast growth factor 2 mRNA. *J. Biol. Chem.*, **280**, 4144–4153.

102 Svitkin, Y.V., Ovchinnikov, L.P., Dreyfuss, G., and Sonenberg, N. (1996) General RNA binding proteins render translation cap dependent. *EMBO J.*, **15**, 7147–7155.

103 LaBranche, H., Dupuis, S., Ben-David, Y., Bani, M.R. *et al.* (1998) Telomere elongation by hnRNP A1 and a derivative that interacts with telomeric repeats and telomerase. *Nat. Genet.*, **19**, 199–202.

104 Zhang, Q.S., Manche, L., Xu, R.M., and Krainer, A.R. (2006) hnRNP A1 associates with telomere ends and stimulates telomerase activity. *RNA*, **12**, 1116–1128.

105 Sawicka, K., Bushell, M., Spriggs, K.A., and Willis, A.E. (2008) Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein. *Biochem. Soc. Trans.*, **36**, 641–647.

106 Garcia-Blanco, M.A., Jamison, S.F., and Sharp, P.A. (1989) Identification and purification of a 62,000-dalton protein that binds specifically to the polypyrimidine tract of introns. *Genes Dev.*, **3**, 1874–1886.

107 Wollerton, M.C., Gooding, C., Wagner, E.J., Garcia-Blanco, M.A. *et al.* (2004) Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. *Mol. Cell*, **13**, 91–100.

108 Castelo-Branco, P., Furger, A., Wollerton, M., Smith, C. *et al.* (2004) Polypyrimidine tract binding protein modulates efficiency of polyadenylation. *Mol. Cell. Biol.*, **24**, 4174–4183.

109 Lou, H., Helfman, D.M., Gagel, R.F., and Berget, S.M. (1999) Polypyrimidine tract-binding protein positively regulates inclusion of an alternative 3′-terminal exon. *Mol. Cell. Biol.*, **19**, 78–85.

110 Jang, S.K. and Wimmer, E. (1990) Cap-independent translation of encephalomyocarditis virus RNA: structural elements of the internal ribosomal

entry site and involvement of a cellular 57-kD RNA-binding protein. *Genes Dev.*, **4**, 1560–1572.

111 Kafasla, P., Lin, H., Curry, S., and Jackson, R.J. (2011) Activation of picornaviral IRESs by PTB shows differential dependence on each PTB RNA-binding domain. *RNA*, **17**, 1120–1131.

112 Hui, J., Hung, L.H., Heiner, M., Schreiner, S. *et al.* (2005) Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. *EMBO J.*, **24**, 1988–1998.

113 Guang, S., Felthauser, A.M., and Mertz, J.E. (2005) Binding of hnRNP L to the pre-mRNA processing enhancer of the herpes simplex virus thymidine kinase gene enhances both polyadenylation and nucleocytoplasmic export of intronless mRNAs. *Mol. Cell. Biol.*, **25**, 6303–6313.

114 Liu, X. and Mertz, J.E. (1995) HnRNP L binds a cis-acting RNA sequence element that enables intron-dependent gene expression. *Genes Dev.*, **9**, 1766–1780.

115 Hamilton, B.J., Nichols, R.C., Tsukamoto, H., Boado, R.J. *et al.* (1999) hnRNP A2 and hnRNP L bind the 3′ UTR of glucose transporter 1 mRNA and exist as a complex in vivo. *Biochem. Biophys. Res. Commun.*, **261**, 646–651.

116 Jafarifar, F., Yao, P., Eswarappa, S.M., and Fox, P.L. (2011) Repression of VEGFA by CA-rich element-binding microRNAs is modulated by hnRNP L. *EMBO J.*, **30**, 1324–1334.

117 Rossbach, O., Hung, L.H., Khrameeva, E., Schreiner, S. *et al.* (2014) Crosslinking-immunoprecipitation (iCLIP) analysis reveals global regulatory roles of hnRNP L. *RNA Biol.*, **11**, 146–155.

118 Ule, J., Stefani, G., Mele, A., Ruggiu, M. *et al.* (2006) An RNA map predicting Nova-dependent splicing regulation. *Nature*, **444**, 580–586.

119 Tange, T.O., Damgaard, C.K., Guth, S., Valcarcel, J. *et al.* (2001) The hnRNP A1 protein regulates HIV-1 tat splicing via a novel intron silencer element. *EMBO J.*, **20**, 5748–5758.

120 Zhou, H.L., Baraniak, A.P., and Lou, H. (2007) Role for Fox-1/Fox-2 in mediating the neuronal pathway of calcitonin/calcitonin gene-related peptide alternative RNA processing. *Mol. Cell. Biol.*, **27**, 830–841.

121 Izquierdo, J.M. (2008) Hu antigen R (HuR) functions as an alternative pre-mRNA splicing regulator of Fas apoptosis-promoting receptor on exon definition. *J. Biol. Chem.*, **283**, 19077–19084.

122 Buratti, E., Baralle, M., De Conti, L., Baralle, D. *et al.* (2004) hnRNP H binding at the 5′ splice site correlates with the pathological effect of two intronic mutations in the NF-1 and TSHbeta genes. *Nucleic Acids Res.*, **32**, 4224–4236.

123 Mauger, D.M., Lin, C., and Garcia-Blanco, M.A. (2008) hnRNP H and hnRNP F complex with Fox2 to silence fibroblast growth factor receptor 2 exon IIIc. *Mol. Cell. Biol.*, **28**, 5403–5419.

124 Carstens, R.P., Wagner, E.J., and Garcia-Blanco, M.A. (2000) An intronic splicing silencer causes skipping of the IIIb exon of fibroblast growth factor receptor 2 through involvement of polypyrimidine tract binding protein. *Mol. Cell. Biol.*, **20**, 7388–7400.

125 Sauliere, J., Sureau, A., Expert-Bezancon, A., and Marie, J. (2006) The polypyrimidine tract binding protein (PTB) represses splicing of exon 6B from the beta-tropomyosin pre-mRNA by directly interfering with the binding of the U2AF65 subunit. *Mol. Cell. Biol.*, **26**, 8755–8769.

126 Blanchette, M. and Chabot, B. (1999) Modulation of exon skipping by high-affinity hnRNP A1-binding sites and by intron elements that repress splice site utilization. *EMBO J.*, **18**, 1939–1952.

127 Martinez-Contreras, R., Fisette, J.F., Nasim, F.U., Madden, R. *et al.* (2006) Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. *PLoS Biol.*, **4**, e21.

128 Hofmann, Y. and Wirth, B. (2002) hnRNP-G promotes exon 7 inclusion of survival motor neuron (SMN) via direct interaction with Htra2-β1. *Hum. Mol. Genet.*, **11**, 2037–2049.

129 Paradis, C., Cloutier, P., Shkreta, L., Toutant, J. *et al.* (2007) hnRNP I/PTB can antagonize the splicing repressor activity of SRp30c. *RNA*, **13**, 1287–1300.

130 Erkelenz, S., Mueller, W.F., Evans, M.S., Busch, A. *et al.* (2013) Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. *RNA*, **19**, 96–102.

131 Grosso, A.R., Gomes, A.Q., Barbosa-Morais, N.L., Caldeira, S. *et al.* (2008) Tissue-specific splicing factor gene expression signatures. *Nucleic Acids Res.*, **36**, 4823–4832.

132 Han, H., Irimia, M., Ross, P.J., Sung, H.K. *et al.* (2013) MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature*, **498**, 241–245.

133 Mallinjoud, P., Villemin, J.P., Mortada, H., Polay Espinoza, M. *et al.* (2014) Endothelial, epithelial, and fibroblast cells exhibit specific splicing programs independently of their tissue of origin. *Genome Res.*, **24**, 511–521.

134 Boukis, L.A., Liu, N., Furuyama, S., and Bruzik, J.P. (2004) Ser/Arg-rich protein-mediated communication between U1 and U2 small nuclear ribonucleoprotein particles. *J. Biol. Chem.*, **279**, 29647–29653.

135 Graveley, B.R., Hertel, K.J., and Maniatis, T. (2001) The role of U2AF35 and U2AF65 in enhancer-dependent splicing. *RNA*, **7**, 806–818.

136 Dembowski, J.A., An, P., Scoulos-Hanson, M., Yeo, G. *et al.* (2012) Alternative splicing of a novel inducible exon diversifies the CASK guanylate kinase domain. *J. Nucleic Acids*, **2012**, 816237.

137 Han, J., Ding, J.H., Byeon, C.W., Kim, J.H. *et al.* (2011) SR proteins induce alternative exon skipping through their activities on the flanking constitutive exons. *Mol. Cell. Biol.*, **31**, 793–802.

138 Zhou, X., Wu, W., Li, H., Cheng, Y. *et al.* (2014) Transcriptome analysis of alternative splicing events regulated by SRSF10 reveals position-dependent splicing modulation. *Nucleic Acids Res.*, **42**, 4019–4030.

139 Blencowe, B.J., Bauren, G., Eldridge, A.G., Issner, R. *et al.* (2000) The SRm160/300 splicing coactivator subunits. *RNA*, **6**, 111–120.

140 Spellman, R. and Smith, C.W. (2006) Novel modes of splicing repression by PTB. *Trends Biochem. Sci*, **31**, 73–76.

141 Izquierdo, J.M., Majos, N., Bonnal, S., Martinez, C. *et al.* (2005) Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. *Mol. Cell*, **19**, 475–484.

142 Sharma, S., Kohlstaedt, L.A., Damianov, A., Rio, D.C. *et al.* (2008) Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. *Nat. Struct. Mol. Biol.*, **15**, 183–191.

143 Sharma, S., Maris, C., Allain, F.H., and Black, D.L. (2011) U1 snRNA directly interacts with polypyrimidine tract-binding protein during splicing repression. *Mol. Cell*, **41**, 579–588.

144 Lamichhane, R., Daubner, G.M., Thomas-Crusells, J., Auweter, S.D. *et al.* (2010) RNA looping by PTB: evidence using FRET and NMR spectroscopy for a role in splicing repression. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 4105–4110.

145 Chiou, N.T., Shankarling, G., and Lynch, K.W. (2013) hnRNP L and hnRNP A1 induce extended U1 snRNA interactions with an exon to repress spliceosome assembly. *Mol. Cell*, **49**, 972–982.

146 Zhu, J., Mayeda, A., and Krainer, A.R. (2001) Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. *Mol. Cell*, **8**, 1351–1361.

147 Chen, C.D., Kobayashi, R., and Helfman, D.M. (1999) Binding of hnRNP H to an exonic splicing silencer is involved in the regulation of alternative splicing of the rat beta-tropomyosin gene. *Genes Dev.*, **13**, 593–606.

148 Chou, M.Y., Rooke, N., Turck, C.W., and Black, D.L. (1999) hnRNP H is a component of a splicing enhancer complex that activates a c-*src* alternative exon in neuronal cells. *Mol. Cell. Biol.*, **19**, 69–77.

149 Barash, Y., Calarco, J.A., Gao, W., Pan, Q. *et al.* (2010) Deciphering the splicing code. *Nature*, **465**, 53–59.

150 Beyer, A.L. and Osheim, Y.N. (1988) Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes Dev.*, **2**, 754–765.

151 Ameur, A., Zaghlool, A., Halvardson, J., Wetterbom, A. *et al.* (2011) Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat. Struct. Mol. Biol.*, **18**, 1435–1440.

152 Bhatt, D.M., Pandya-Jones, A., Tong, A.J., Barozzi, I. *et al.* (2012) Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell*, **150**, 279–290.

153 Carrillo Oesterreich, F., Preibisch, S., and Neugebauer, K.M. (2010) Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. *Mol. Cell*, **40**, 571–581.

154 Girard, C., Will, C.L., Peng, J., Makarov, E.M. *et al.* (2012) Post-transcriptional spliceosomes are retained in nuclear speckles until splicing completion. *Nat. Commun.*, **3**, 994.

155 Khodor, Y.L., Menet, J.S., Tolan, M., and Rosbash, M. (2012) Cotranscriptional splicing efficiency differs dramatically between *Drosophila* and mouse. *RNA*, **18**, 2174–2186.

156 Khodor, Y.L., Rodriguez, J., Abruzzi, K.C., Tang, C.H. *et al.* (2011) Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in *Drosophila*. *Genes Dev.*, **25**, 2502–2512.

157 Windhager, L., Bonfert, T., Burger, K., Ruzsics, Z. *et al.* (2012) Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome Res.*, **22**, 2031–2042.

158 Brugiolo, M., Herzel, L., and Neugebauer, K.M. (2013) Counting on co-transcriptional splicing. *F1000Prime Rep.*, **5**, 9.

159 Cramer, P., Pesce, C.G., Baralle, F.E., and Kornblihtt, A.R. (1997) Functional association between promoter structure and transcript alternative splicing. *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 11456–11460.

160 Pagani, F., Stuani, C., Zuccato, E., Kornblihtt, A.R. *et al.* (2003) Promoter architecture modulates CFTR exon 9 skipping. *J. Biol. Chem.*, **278**, 1511–1517.

161 Kadener, S., Cramer, P., Nogues, G., Cazalla, D. *et al.* (2001) Antagonistic effects of T-Ag and VP16 reveal a role for RNA pol II elongation on alternative splicing. *EMBO J.*, **20**, 5759–5768.

162 Nogues, G., Kadener, S., Cramer, P., Bentley, D. *et al.* (2002) Transcriptional activators differ in their abilities to control alternative splicing. *J. Biol. Chem.*, **277**, 43110–43114.

163 Auboeuf, D., Dowhan, D.H., Kang, Y.K., Larkin, K. *et al.* (2004) Differential recruitment of nuclear receptor coactivators may determine alternative RNA splice site choice in target genes. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 2270–2274.

164 Auboeuf, D., Dowhan, D.H., Li, X., Larkin, K. *et al.* (2004) CoAA, a nuclear receptor coactivator protein at the interface of transcriptional coactivation and RNA splicing. *Mol. Cell. Biol.*, **24**, 442–453.

165 Auboeuf, D., Honig, A., Berget, S.M., and O'Malley, B.W. (2002) Coordinate regulation of transcription and splicing by steroid receptor coregulators. *Science*, **298**, 416–419.

166 Kadener, S., Fededa, J.P., Rosbash, M., and Kornblihtt, A.R. (2002) Regulation of alternative splicing by a transcriptional enhancer through RNA pol II elongation. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 8185–8190.

167 Batsche, E., Yaniv, M., and Muchardt, C. (2006) The human SWI/SNF subunit Brm is a regulator of alternative splicing. *Nat. Struct. Mol. Biol.*, **13**, 22–29.

168 Allo, M., Buggiano, V., Fededa, J.P., Petrillo, E. *et al.* (2009) Control of alternative splicing through siRNA-mediated transcriptional gene silencing. *Nat. Struct. Mol. Biol.*, **16**, 717–724.

169 Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J. *et al.* (2010) Regulation of alternative splicing by histone modifications. *Science*, **327**, 996–1000.

170 Pradeepa, M.M., Sutherland, H.G., Ule, J., Grimes, G.R. *et al.* (2012) Psip1/Ledgf p52 binds methylated histone H3K36 and splicing factors and contributes to the regulation of alternative splicing. *PLos Genet.*, **8**, e1002717.

171 Saint-Andre, V., Batsche, E., Rachez, C., and Muchardt, C. (2011) Histone H3 lysine 9 trimethylation and HP1gamma favor inclusion of alternative exons. *Nat. Struct. Mol. Biol.*, **18**, 337–344.

172 Schor, I.E., Rascovan, N., Pelisch, F., Allo, M. *et al.* (2009) Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 4325–4330.

173 Kornblihtt, A.R., Schor, I.E., Allo, M., Dujardin, G. *et al.* (2013) Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat. Rev. Mol. Cell Biol.*, **14**, 153–165.

174 de la Mata, M. and Kornblihtt, A.R. (2006) RNA polymerase II C-terminal domain mediates regulation of alternative splicing by SRp20. *Nat. Struct. Mol. Biol.*, **13**, 973–980.

175 Shukla, S., Kavak, E., Gregory, M., Imashimizu, M. *et al.* (2011) CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, **479**, 74–79.

176 Schwartz, S., Meshorer, E., and Ast, G. (2009) Chromatin organization marks exon-intron structure. *Nat. Struct. Mol. Biol.*, **16**, 990–995.

177 Spies, N., Nielsen, C.B., Padgett, R.A., and Burge, C.B. (2009) Biased chromatin signatures around polyadenylation sites and exons. *Mol. Cell*, **36**, 245–254.

178 Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M. *et al.* (2009) Nucleosome positioning as a determinant of exon recognition. *Nat. Struct. Mol. Biol.*, **16**, 996–1001.

179 Zhou, H.L., Luo, G., Wise, J.A., and Lou, H. (2014) Regulation of alternative splicing by local histone modifications: potential roles for RNA-guided mechanisms. *Nucleic Acids Res.*, **42**, 701–713.

180 Nicholson, P., Yepiskoposyan, H., Metze, S., Zamudio Orozco, R. *et al.* (2010) Nonsense-mediated mRNA decay in human cells: mechanistic insights, functions beyond quality control and the double-life of NMD factors. *Cell. Mol. Life Sci.*, **67**, 677–700.

181 Hillman, R.T., Green, R.E., and Brenner, S.E. (2004) An unappreciated role for RNA surveillance. *Genome Biol.*, **5**, R8.

182 Nagy, E. and Maquat, L.E. (1998) A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem. Sci*, **23**, 198–199.

183 Green, R.E., Lewis, B.P., Hillman, R.T., Blanchette, M. *et al.* (2003) Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes. *Bioinformatics*, **19** (Suppl 1), i118–i121.

184 Mendell, J.T., Sharifi, N.A., Meyers, J.L., Martinez-Murillo, F. *et al.* (2004) Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. *Nat. Genet.*, **36**, 1073–1078.

185 Wittmann, J., Hol, E.M., and Jack, H.M. (2006) hUPF2 silencing identifies physiologic substrates of mammalian nonsense-mediated mRNA decay. *Mol. Cell. Biol.*, **26**, 1272–1287.

186 Pan, Q., Saltzman, A.L., Kim, Y.K., Misquitta, C. *et al.* (2006) Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. *Genes Dev.*, **20**, 153–158.

187 Rossbach, O., Hung, L.H., Schreiner, S., Grishina, I. *et al.* (2009) Auto- and cross-regulation of the hnRNP L proteins by alternative splicing. *Mol. Cell. Biol.*, **29**, 1442–1451.

188 Saltzman, A.L., Kim, Y.K., Pan, Q., Fagnani, M.M. *et al.* (2008) Regulation of multiple core spliceosomal proteins by alternative splicing-coupled nonsense-mediated mRNA decay. *Mol. Cell. Biol.*, **28**, 4320–4330.

189 Lykke-Andersen, S. and Jensen, T.H. (2015) Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat. Rev. Mol. Cell Biol.*, **16**, 665–677.

190 Li, T., Shi, Y., Wang, P., Guachalla, L.M. *et al.* (2015) Smg6/Est1 licenses embryonic stem cell differentiation via nonsense-mediated mRNA decay. *EMBO J.*, **34**, 1630–1647.

191 McIlwain, D.R., Pan, Q., Reilly, P.T., Elia, A.J. *et al.* (2010) Smg1 is required for embryogenesis and regulates diverse genes via alternative splicing coupled to nonsense-mediated mRNA decay. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 12186–12191.

192 Medghalchi, S.M., Frischmeyer, P.A., Mendell, J.T., Kelly, A.G. *et al.* (2001) Rent1, a trans-effector of nonsense-mediated mRNA decay, is essential for Mammalian embryonic viability. *Hum. Mol. Genet.*, **10**, 99–105.

193 Weischenfeldt, J., Damgaard, I., Bryder, D., Theilgaard-Monch, K. *et al.* (2008) NMD is essential for hematopoietic stem and progenitor cells and for eliminating by-products of programmed DNA rearrangements. *Genes Dev.*, **22**, 1381–1396.

194 David, C.J. and Manley, J.L. (2010) Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes Dev.*, **24**, 2343–2364.

195 Mills, J.D. and Janitz, M. (2012) Alternative splicing of mRNA in the molecular pathology of neurodegenerative diseases. *Neurobiol. Aging*, **33** (1012), e11–e24.

196 Singh, R.K. and Cooper, T.A. (2012) Pre-mRNA splicing in disease and therapeutics. *Trends Mol. Med.*, **18**, 472–482.

197 Krawczak, M., Thomas, N.S., Hundrieser, B., Mort, M. *et al.* (2007) Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. *Hum. Mutat.*, **28**, 150–158.

198 Lim, K.H., Ferraris, L., Filloux, M.E., Raphael, B.J. *et al.* (2011) Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 11093–11098.

199 Buratti, E., Chivers, M., Hwang, G., and Vorechovsky, I. (2011) DBASS3 and DBASS5: databases of aberrant 3′- and 5′-splice sites. *Nucleic Acids Res.*, **39**, D86–D91.

200 Fukumaki, Y., Ghosh, P.K., Benz, E.J. Jr., Reddy, V.B. *et al.* (1982) Abnormally spliced messenger RNA in erythroid cells from patients with beta+ thalassemia and monkey cells expressing a cloned beta+-thalassemic gene. *Cell*, **28**, 585–593.

201 Law, A.J., Kleinman, J.E., Weinberger, D.R., and Weickert, C.S. (2007) Disease-associated intronic variants in the ErbB4 gene are related to altered ErbB4 splice-variant expression in the brain in schizophrenia. *Hum. Mol. Genet.*, **16**, 129–141.

202 Quesada, V., Conde, L., Villamor, N., Ordonez, G.R. *et al.* (2012) Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat. Genet.*, **44**, 47–52.

203 Arnold, E.S., Ling, S.C., Huelga, S.C., Lagier-Tourenne, C. *et al.* (2013) ALS-linked TDP-43 mutations produce aberrant RNA splicing and adult-onset motor neuron disease without aggregation or loss of nuclear TDP-43. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, E736–E745.

204 Lagier-Tourenne, C., Polymenidou, M., and Cleveland, D.W. (2010) TDP-43 and FUS/TLS: emerging roles in RNA processing and neurodegeneration. *Hum. Mol. Genet.*, **19**, R46–R64.

205 Spitali, P. and Aartsma-Rus, A. (2012) Splice modulating therapies for human disease. *Cell*, **148**, 1085–1088.

206 Lorson, C.L., Hahnen, E., Androphy, E.J., and Wirth, B. (1999) A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 6307–6311.

207 Hua, Y., Sahashi, K., Rigo, F., Hung, G. *et al.* (2011) Peripheral SMN restoration is essential for long-term rescue of a severe spinal muscular atrophy mouse model. *Nature*, **478**, 123–126.

208 Ottesen, E.W. (2017) ISS-N1 makes the first FDA-approved drug for spinal muscular atrophy. *Transl. Neurosci.*, **8**, 1–6.

209 Grund, S.E., Fischer, T., Cabal, G.G., Antunez, O. *et al.* (2008) The inner nuclear membrane protein Src1 associates with subtelomeric genes and alters their regulated gene expression. *J. Cell Biol.*, **182**, 897–910.

210 Rodriguez-Navarro, S., Igual, J.C., and Perez-Ortin, J.E. (2002) SRC1: an intron-containing yeast gene involved in sister chromatid segregation. *Yeast*, **19**, 43–54.

211 Kress, T.L., Krogan, N.J., and Guthrie, C. (2008) A single SR-like protein, Npl3, promotes pre-mRNA splicing in budding yeast. *Mol. Cell*, **32**, 727–734.

212 Spingola, M. and Ares, M. Jr. (2000) A yeast intronic splicing enhancer and Nam8p are required for Mer1p-activated splicing. *Mol. Cell*, **6**, 329–338.

213 Dong, S., Li, C., Zenklusen, D., Singer, R.H. *et al.* (2007) YRA1 autoregulation requires nuclear export and cytoplasmic Edc3p-mediated degradation of its pre-mRNA. *Mol. Cell*, **25**, 559–573.

214 Preker, P.J. and Guthrie, C. (2006) Autoregulation of the mRNA export factor Yra1p requires inefficient splicing of its pre-mRNA. *RNA*, **12**, 994–1006.

215 Pleiss, J.A., Whitworth, G.B., Bergkessel, M., and Guthrie, C. (2007) Transcript specificity in yeast pre-mRNA splicing revealed by mutations in core spliceosomal components. *PLoS Biol.*, **5**, e90.

216 Sayani, S., Janis, M., Lee, C.Y., Toesca, I. *et al.* (2008) Widespread impact of nonsense-mediated mRNA decay on the yeast intronome. *Mol. Cell*, **31**, 360–370.

217 Juneau, K., Palm, C., Miranda, M., and Davis, R.W. (2007) High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 1522–1527.

218 Munding, E.M., Shiue, L., Katzman, S., Donohue, J.P. *et al.* (2013) Competition between pre-mRNAs for the splicing machinery drives global regulation of splicing. *Mol. Cell*, **51**, 338–348.

219 Parenteau, J., Durand, M., Morin, G., Gagnon, J. *et al.* (2011) Introns within ribosomal protein genes regulate the production and function of yeast ribosomes. *Cell*, **147**, 320–331.

220 Parenteau, J., Durand, M., Veronneau, S., Lacombe, A.A. *et al.* (2008) Deletion of many yeast introns reveals a minority of genes that require splicing for function. *Mol. Biol. Cell*, **19**, 1932–1941.

221 Volanakis, A., Passoni, M., Hector, R.D., Shah, S. *et al.* (2013) Spliceosome-mediated decay (SMD) regulates expression of nonintronic genes in budding yeast. *Genes Dev.*, **27**, 2025–2038.

222 Gould, G.M., Paggi, J.M., Guo, Y., Phizicky, D.V. *et al.* (2016) Identification of new branch points and unconventional introns in *Saccharomyces cerevisiae*. *RNA*, **22**, 1522–1534.

223 Kawashima, T., Douglass, S., Gabunilas, J., Pellegrini, M. *et al.* (2014) Widespread use of non-productive alternative splice sites in *Saccharomyces cerevisiae*. *PLos Genet.*, **10**, e1004249.

224 Schreiber, K., Csaba, G., Haslbeck, M., and Zimmer, R. (2015) Alternative splicing in next generation sequencing data of *Saccharomyces cerevisiae*. *PLoS One*, **10**, e0140487.

225 Roth, A. and Breaker, R.R. (2009) The structural and functional diversity of metabolite-binding riboswitches. *Annu. Rev. Biochem.*, **78**, 305–334.

226 Ramesh, A. (2015) Second messenger – sensing riboswitches in bacteria. *Semin. Cell Dev. Biol.*, **47–48**, 3–8.

227 Serganov, A. and Nudler, E. (2013) A decade of riboswitches. *Cell*, **152**, 17–24.

228 Lee, E.R., Baker, J.L., Weinberg, Z., Sudarsan, N. *et al.* (2010) An allosteric self-splicing ribozyme triggered by a bacterial second messenger. *Science*, **329**, 845–848.

229 Sudarsan, N., Lee, E.R., Weinberg, Z., Moy, R.H. *et al.* (2008) Riboswitches in eubacteria sense the second messenger cyclic di-GMP. *Science*, **321**, 411–413.

230 Chen, A.G., Sudarsan, N., and Breaker, R.R. (2011) Mechanism for gene control by a natural allosteric group I ribozyme. *RNA*, **17**, 1967–1972.

231 Sudarsan, N., Barrick, J.E., and Breaker, R.R. (2003) Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA*, **9**, 644–647.

232 Wachter, A. (2010) Riboswitch-mediated control of gene expression in eukaryotes. *RNA Biol.*, **7**, 67–76.

233 Kubodera, T., Watanabe, M., Yoshiuchi, K., Yamashita, N. *et al.* (2003) Thiamine-regulated gene expression of *Aspergillus oryzae thiA* requires splicing of the intron containing a riboswitch-like domain in the 5′-UTR. *FEBS Lett.*, **555**, 516–520.

234 Cheah, M.T., Wachter, A., Sudarsan, N., and Breaker, R.R. (2007) Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. *Nature*, **447**, 497–500.

235 Li, S. and Breaker, R.R. (2013) Eukaryotic TPP riboswitch regulation of alternative splicing involving long-distance base pairing. *Nucleic Acids Res.*, **41**, 3022–3031.

236 Bocobza, S., Adato, A., Mandel, T., Shapira, M. *et al.* (2007) Riboswitch-dependent gene regulation and its evolution in the plant kingdom. *Genes Dev.*, **21**, 2874–2879.

237 Wachter, A., Tunc-Ozdemir, M., Grove, B.C., Green, P.J. *et al.* (2007) Riboswitch control of gene expression in plants by splicing and alternative 3′ end processing of mRNAs. *Plant Cell*, **19**, 3437–3450.

238 Borsuk, P., Przykorska, A., Blachnio, K., Koper, M. *et al.* (2007) l-arginine influences the structure and function of arginase mRNA in *Aspergillus nidulans*. *Biol. Chem.*, **388**, 135–144.

239 Le Hir, H., Nott, A., and Moore, M.J. (2003) How introns influence and enhance eukaryotic gene expression. *Trends Biochem. Sci*, **28**, 215–220.

240 Hamer, D.H. and Leder, P. (1979) Splicing and the formation of stable RNA. *Cell*, **18**, 1299–1302.

241 Buchman, A.R. and Berg, P. (1988) Comparison of intron-dependent and intron-independent gene expression. *Mol. Cell. Biol.*, **8**, 4395–4405.

242 Choi, T., Huang, M., Gorman, C., and Jaenisch, R. (1991) A generic intron increases gene expression in transgenic mice. *Mol. Cell. Biol.*, **11**, 3070–3074.
243 Luehrsen, K.R. and Walbot, V. (1991) Intron enhancement of gene expression and the splicing efficiency of introns in maize cells. *Mol. Gen. Genet.*, **225**, 81–93.
244 Nott, A., Meislin, S.H., and Moore, M.J. (2003) A quantitative analysis of intron effects on mammalian gene expression. *RNA*, **9**, 607–617.
245 Zieler, H. and Huynh, C.Q. (2002) Intron-dependent stimulation of marker gene expression in cultured insect cells. *Insect Mol. Biol.*, **11**, 87–95.
246 Bourdon, V., Harvey, A., and Lonsdale, D.M. (2001) Introns and their positions affect the translational activity of mRNA in plant cells. *EMBO Rep.*, **2**, 394–398.
247 Hodges, D. and Crooke, S.T. (1995) Inhibition of splicing of wild-type and mutated luciferase-adenovirus pre-mRNAs by antisense oligonucleotides. *Mol. Pharmacol.*, **48**, 905–918.
248 Lacy-Hulbert, A., Thomas, R., Li, X.P., Lilley, C.E. *et al.* (2001) Interruption of coding sequences by heterologous introns can enhance the functional expression of recombinant genes. *Gene Ther.*, **8**, 649–653.
249 Huang, M.T. and Gorman, C.M. (1990) Intervening sequences increase efficiency of RNA 3′ processing and accumulation of cytoplasmic RNA. *Nucleic Acids Res.*, **18**, 937–947.
250 Gusti, V., Kim, D.S., and Gaur, R.K. (2008) Sequestering of the 3′ splice site in a theophylline-responsive riboswitch allows ligand-dependent control of alternative splicing. *Oligonucleotides*, **18**, 93–99.
251 Kim, D.S., Gusti, V., Pillai, S.G., and Gaur, R.K. (2005) An artificial riboswitch for controlling pre-mRNA splicing. *RNA*, **11**, 1667–1677.
252 Kim, D.S., Gusti, V., Dery, K.J., and Gaur, R.K. (2008) Ligand-induced sequestering of branchpoint sequence allows conditional control of splicing. *BMC Mol. Biol.*, **9**, 23.
253 Weigand, J.E. and Suess, B. (2007) Tetracycline aptamer-controlled regulation of pre-mRNA splicing in yeast. *Nucleic Acids Res.*, **35**, 4179–4185.
254 Culler, S.J., Hoff, K.G., and Smolke, C.D. (2010) Reprogramming cellular behavior with RNA controllers responsive to endogenous proteins. *Science*, **330**, 1251–1255.
255 Hickey, S.F., Sridhar, M., Westermann, A.J., Qin, Q. *et al.* (2012) Transgene regulation in plants by alternative splicing of a suicide exon. *Nucleic Acids Res.*, **40**, 4701–4710.

8

Design of Ligand-Controlled Genetic Switches Based on RNA Interference

Shunnichi Kashida[1,2] and Hirohide Saito[1]

[1] *Center for iPS Cell Research and Application, Kyoto University, Department of Life Science Frontiers, 53 Kawahara-cho, Shogoin, Sakyo-ku, Kyoto 606-8507, Japan*
[2] *Ecole Normale Supérieure, UMR 8640 CNRS-ENS-UPMC Pasteur, Department of Chemistry, 24 rue Lhomond Paris, 75005, France*

8.1 Utility of the RNAi Pathway for Application in Mammalian Cells

RNA interference (RNAi) is an efficient and convenient tool for transient gene suppression (knockdown) in biomedical research. RNAi is beneficial for genetic screening and basic studies involving loss-of-function phenotypes and as an alternative protein inhibitor to small molecule drugs [1]. Since the first discovery of the RNAi phenomena in *Caenorhabditis elegans* [2], intensive genetic and biochemical research has uncovered the molecular mechanisms underlying RNAi and identified analogous pathways and molecules to control RNAi in eukaryotes [3–5].

In mammalian systems, RNAi is induced when microRNA (miRNA), short hairpin RNA (shRNA), or small interfering RNA (siRNA) harness the endogenous processing pathway and machinery (Figure 8.1). In the endogenous RNAi pathway, primary miRNA (pri-miRNA) embedded in coding or noncoding RNA is transcribed from genetic or plasmid DNA by RNA polymerase II or III and is cleaved at the base region of the stem-loop structure (two black wedges) by the RNase III nuclease Drosha. The cleaved stem-loop precursor miRNA (pre-miRNA) is recognized by Expotin-5b proteins, exported from the nucleus to the cytoplasm and processed into mature miRNA (two black wedges) by another RNase III nuclease, Dicer. Then, one strand of the mature miRNA is selected and introduced into Ago2 to activate sequence-specific mRNA degradation and targeted gene repression. shRNA expressed from plasmids is exported to the cytoplasm and processed only by Dicer in a similar manner to transfected shRNA or siRNA molecules (Figure 8.1).

From a synthetic biology perspective, RNAi is a suitable and potent technology for the development of genetic devices to rewire cell signaling. It is important to generate RNAi-modulated genetic devices that detect target input molecules

Synthetic Biology: Parts, Devices and Applications, First Edition. Edited by Christina Smolke.

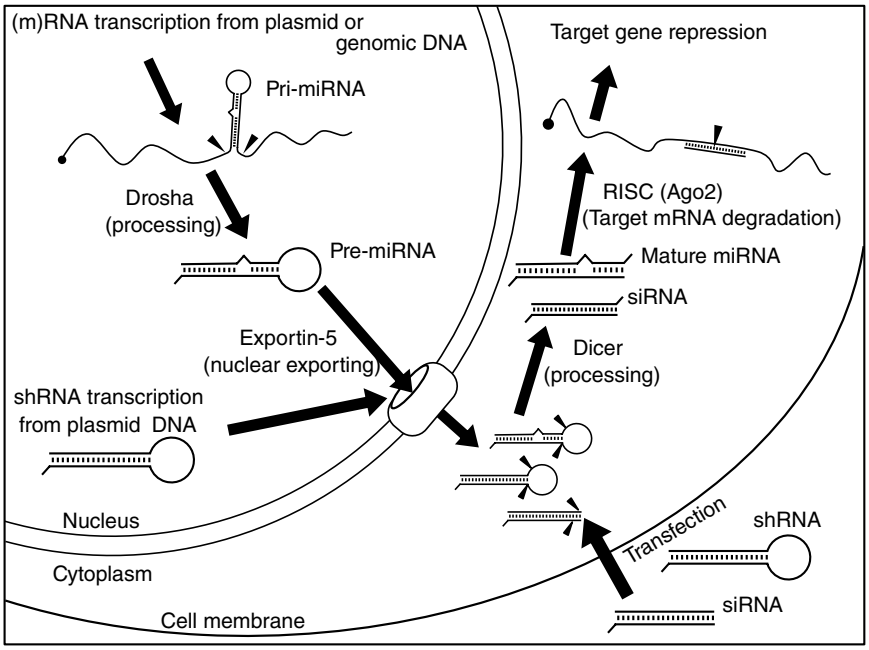


Figure 8.1 Schematic of the RNAi pathway in mammalian cells, including endogenous miRNA processing and ectopic shRNA or siRNA expression.

(triggers) and control RNAi-mediated gene expression (referred to as RNAi switches). To obtain such RNAi switches, appropriate RNA sequences located in (pri)(pre)-miRNAs or shRNAs have been engineered to modulate the recognition of RNA-processing nucleases such as Drosha and Dicer, making it possible to bind to various triggers and inhibit or permit nuclease processing. To design the trigger molecule-controlled RNAi switches, it is useful to isolate functional RNA modules based on RNA secondary structures because RNA is often divided into functional modules and reassembled through the double-stranded regions without disrupting the original function. Synthetic RNAi switches have been developed by employing various trigger molecules (e.g., small molecules, RNA, or proteins) that take advantage of the modularity of RNA.

8.2 Development of RNAi Switches that Respond to Trigger Molecules

Control of gene expression from exogenous DNA by a set of transcription factors and coupled small molecules has conventionally been used for conditional expression strategies [6–8]. Similarly, the transcriptional control of shRNA expression using small molecules has been employed for the construction of tunable genetic switches based on RNAi [9]. This system has also combined the Lac inhibitor with shRNA RNAi to synergistically suppress target gene transcription and translation.

Hereafter, we will focus on an RNA design strategy and the posttranscriptional gene expression control of RNAi switches via several trigger molecules including small molecules, oligonucleotides, and proteins (Figure 8.2). These triggers bind to specific RNA sequences, and the interactions between them can be employed to generate RNAi switches (Table 8.1). The advantages and potential applications of RNAi switches primarily depend on the type of trigger. Small molecule triggers that penetrate through the cell membrane tune the function of RNAi switches by adjusting the extracellular concentration of the input molecules, which is a mechanism similar to that of small molecule-inducible transcription factors. Oligonucleotide triggers, such as DNA, RNA, and modified oligonucleotides (MONs), are able to form Watson–Crick base pairs with designed RNA devices and thus adjust specificities and affinities between the trigger molecules and the devices. Protein triggers can also be used to control the functions of RNAi switches. Thus, specific proteins expressed in cells can distinguish target cell types based on the intracellular environment.

8.2.1 Small Molecule-Triggered RNAi Switches

Small molecule-triggered RNAi switches have been designed to modulate Dicer or Drosha processing of shRNA or pri-miRNA. Initially, three different switch design strategies implementing a theophylline aptamer were employed to achieve theophylline-responsive properties [20]. In the first design approach to obtain theophylline-responsive shRNA switches, the loop region of EGFP- (or DsRed-) targeting shRNA was replaced with a theophylline aptamer containing a loop sequence; this replacement was designed to create a theophylline and RNA complex around the Dicer recognition site [10]. When expressed in HEK293 cells, the switches inhibited Dicer processing and knockdown of reporter fluorescent genes in the presence of theophylline in culture medium. A similar approach was applied to the development of pri-miRNA-based RNAi switches (pri-miRNA

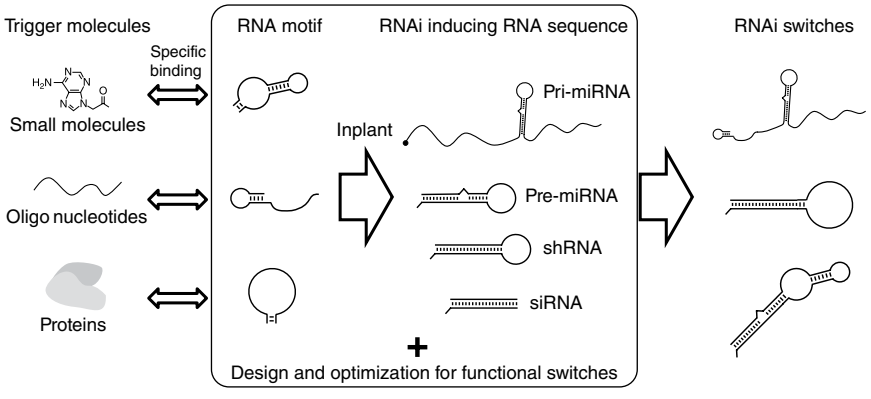


Figure 8.2 RNAi switch design strategies with a variety of trigger molecules. The RNA motifs that bind to specific trigger molecules are introduced into the appropriate regions in pri-miRNA, pre-miRNA, shRNA, or siRNA. The motifs embedded in the RNA are then optimized to generate functional RNAi switches.

Table 8.1 Developed RNAi switches.

References	An *et al.* [10]	Beisel *et al.* [11]	Kumar *et al.* [12]	Masu *et al.* [13]	Beisel *et al.* [14]	Kumar *et al.* [15]	Saito *et al.* [16]	Kashida *et al.* [17]	Afonin *et al.* [18]	Velagapudi *et al.* [19]
Input	Small molecule (theophylline)	Small molecule (theophylline, hypoxanthine)	Small molecule (theophylline)	RNA	Small molecule (theophylline, hypoxanthine)	Chemically modified RNA	Archaeal protein (L7Ae)	Human protein (U1A, NFκB)	Chemically modified DNA	Small molecule library
Dose	~10 mM	~5 mM	~3 mM	~500 nM	~5 mM	~500 nM	Plasmid expression	Plasmid expression	~500 nM	~40 μM
Based RNA	shRNA	shRNA	Pri-miRNA	siRNA	Pri-miRNA	Pri-miRNA + siRNA	shRNA	shRNA	siRNA	Pre-miRNA
Inhibiting machinery	Dicer	Dicer	Drosha	Dicer	Drosha	Drosha + Dicer	Dicer	Dicer	Dicer	Drosha
Switch	**ON** → OFF	**ON** → OFF	OFF → **ON**	OFF → **ON**	**ON** → OFF	OFF → **ON**	**ON** → OFF	**ON** → OFF	OFF → **ON**	ON → **OFF**
Target gene	EGFP, DsRed	EGFP	EGFP	Luciferase	EGFP	EGFP, DsRed	EGFP, antiapoptotic gene (Bcl-xL)	EGFP	EGFP, HIV-1 Gag-pol glycoprotein	Proapoptotic gene (FOXO1)
Cell type	HEK 293	HEK 293-EGFP	HEK 293	HeLa	HEK293(T), HeLa, MDA-MB-231, HEK293T-EGFP	HEK 293	HeLa, HeLa-EGFP	HEK293FT	HeLa, MDA-MB231, MDA-MB231-EGFP	MCF7

switches), and the theophylline aptamer was introduced into the Drosha recognition site of pri-miRNA [14]. The second strategy to design switches was based on the changing stability of the two RNA reversible conformational states (active or inactive) via trigger (theophylline or hypoxanthine) binding [11]. In the absence of triggers, shRNA switched from the canonical dsRNA structure (active state) that is required for EGFP knockdown; the binding of the trigger changed the secondary structure of the switches, and part of one dsRNA strand stably bound to the adjacent loop sequence (inactive state) and collapsed the canonical dsRNA structure. The third strategy employed an irreversible conformational change of the pri-miRNA structure and ligand-controlled hammerhead ribozymes [12]. In the absence of theophylline, the allosteric hammerhead ribozyme domain and the following inhibitory strand that hybridizes the pri-miRNA collapsed the canonical structure that is required for Drosha processing. In the presence of theophylline, ribozyme self-cleavage induced the exposure of the 5′-single-stranded region that was originally masked by the inhibitory strand, resulting in Drosha processing. Another design strategy considered endogenous pre-miRNA as potential RNAi switches. In this system, Dicer or Drosha processing was inhibited by bioactive small molecules that target pre-miRNAs [19]. The strategy employed the tight binding pair of a benzimidazole and RNA internal loop motif from a database of RNA motif-small molecule interactions and searched the Dicer or Drosha recognition sites of disease-related pre-miRNA for the RNA internal loop motif [19]. The motif was found and well fit with the Drosha recognition site of human pre-miR-96. In the result the benzimidazole specifically inhibited the endogenous pre-miR-96 maturation, recovered the downstream proapoptotic gene FOXO1 expression and induced apoptosis in MCF7 cells.

8.2.2 Oligonucleotide-Triggered RNAi Switches

Oligonucleotide-triggered RNAi switches have been designed to modulate Dicer processing of siRNA or Drosha processing of pri-miRNA. The strategy for designing the switches is based on toehold-mediated oligonucleotide displacement. DNA-mediated siRNA switches are composed of two DNA–RNA hybrids [18]. The RNA strands of the hybrids are split sense and antisense strands of siRNA. The DNA strands contain additional nucleotides (toeholds) in the hybridization region, and the two DNA strands can potentially form a double strand. The initial DNA–RNA hybrid is not processed by Dicer because it cannot recognize a DNA–RNA hybrid [21]. After the DNA–RNA hybrids are transfected into mammalian MDA-MB231 cells, the DNAs bind to each other at the toehold region and replace RNA with antisense DNA to produce double-stranded DNA and siRNA. The siRNA is then processed by Dicer to knock down target genes.

A small RNA-triggered siRNA switch was designed to block the double-stranded formation of sense strand and antisense strand RNA in the absence of a trigger. For this, an inhibitory sequence is connected to the 3′ end of the sense strand and is partially hybridized with the antisense strand-binding region of the sense strand [13]. Meanwhile, the trigger RNA when present hybridizes with an

inhibitory sequence, resulting in perfect hybridization of the sense and antisense strands. The resulting dsRNA is processed as active siRNA by Dicer (OFF-to-ON RNAi switch). MON-triggered pri-miRNA switches were also designed by introducing an inhibitory RNA stem loop into the 3′ end of the pri-miRNA to conceal the single-stranded region of the Drosha recognition site [15]. MON targets endogenous genes, and the pri-miRNA contains a target sequence for a fluorescent protein (EGFP or DsRed) in EGFP- or DsRed-expressing HeLa cells. When present, MON perfectly hybridizes with half of the inhibitory RNA stem-loop sequence, resulting in an RNA conformational change and exposure of the single-stranded region that is recognized by Drosha. The resulting dsRNA is also processed as an siRNA by Dicer.

8.2.3 Protein-Triggered RNAi Switches

Protein-triggered RNA switches have been designed by replacing the loop region of shRNA with protein-binding sequences in an attempt to mask the Dicer recognition site in the presence of trigger protein molecules [16, 17]. The specific and tight RNA–protein interaction (RNP) motif is important when designing an efficient RNAi switch. For these switches, an RNP motif consisting of an archaeal ribosomal protein, L7Ae, and its binding partner, box C/D kink-turn RNA (Kt), is employed to develop L7Ae-triggered shRNA switches (Kt-shRNA) [16]. L7Ae binds to the loop region of Kt-shRNA, which inhibits Dicer cleavage and targets gene knockdown. Following Kt-shRNA development, designed shRNA switches triggered by the human splicing-related protein U1A and the human transcriptional regulator NFκB (p50 domain) were developed [17]. The RNP motifs of the U1A protein and the loop sequence of U1 snRNA or the loop-stem-loop sequence in the 3′ untranslated region of U1A mRNA were utilized to develop two types of U1A-triggered shRNA switches. An RNP motif composed of the NFκB protein and an artificially selected NFκB-binding aptamer was also employed to develop an NFκB-triggered shRNA switch. The molecular structures of these RNP motifs were solved via crystal or nuclear magnetic resonance (NMR) structural analyses and utilized to create three-dimensional (3D) molecular designs of the switches. The switches were designed by incorporating a protein-binding sequence into the loop region of shRNA, which contains 22–28 bp of dsRNA targeting the EGFP gene in the stem region. The configurations of these switches were three-dimensionally optimized such that the interaction between the trigger protein and shRNA efficiently blocked Dicer processing.

8.3 Rational Design of Functional RNAi Switches

Rational and predictable RNA design strategies are critical for developing versatile RNAi switch systems. Hereafter, we will focus on design strategies for RNAi switches. The most common design strategy for RNA switches utilizes predicted RNA secondary structures and their free energies based on Watson–Crick base pairing in the presence/absence of trigger molecules. This strategy also attempts to optimize the free energy difference between the two states by changing base

pair lengths and introducing mutations [11, 15, 22–25]. Several ligand- (e.g., small molecule or oligonucleotide) responsive RNAi switches have been designed based on this strategy. When the secondary structure, free energy difference, and base pair length have been optimized, a similar design strategy could be applied to generate various RNAi switches that respond to different trigger molecules.

A useful and efficient 3D design approach has been utilized to develop protein-triggered RNAi switches by employing available 3D RNP structures (analyzed via both NMR and X-ray crystallography) [17, 26]. For the first approach, the structural components of shRNA switches were three-dimensionally reconstructed *in silico* by creating 22–28 bp of A-form dsRNA with 3D molecular design software and loading the RNP motif, composed of the trigger protein and its binding RNA motif, from the Protein Data Bank (Figure 8.3a, left). Then, 3D structural models of the trigger protein-bound shRNA switches were constructed by superimposing the few terminal nucleotides of the RNA loop on the dsRNA using minimization methods consisting of the least squares approximation polynomial and connecting the loop with the dsRNA. The models predicted the structural states of the shRNA switches in the presence of the trigger protein (Figure 8.3a, right). As described in Figure 8.3, the bound trigger protein on the shRNA switch rotates approximately 30° in a counterclockwise direction around the axis of the dsRNA with a 1-bp insertion and is located ~2.6 Å farther from the site of Dicer cleavage.

Because the Dicer enzyme can access the 22nd nucleotides from both the 5′ and 3′ ends [27], the bound trigger protein on the switches was designed to block Dicer access. Specifically, the base pair lengths of the switches were adjusted by taking advantage of the orientation change of each base pair such that the bound trigger proteins could block Dicer access (Figure 8.3b). To predict *in silico* the collision between Dicer and the bound trigger protein, the constructed switch models were superimposed on the catalytic sites and the peripheral region of *Giardia* Dicer with reference to the Dicer cleavage sites and catalytic sites. Based on the results of the 3D molecular design and switch assessment, steric hindrance between Dicer and the shRNA-bound protein was predicted *in silico*, which positively correlated with the inhibition of Dicer cleavage *in vitro* and target gene expression in living cells. Furthermore, the 3D molecular design method could be applied for all switches that sense several different RNA-binding proteins (e.g., L7Ae, U1A, and NFκB) and could be used to predict the functions of these proteins. In principle, the strategy could predict functional switch structures in response to RNA-binding proteins to adjust the ON/OFF ratio of the designed switches.

8.4 Application of the RNAi Switches

RNAi switches have been proposed for applications including drug delivery, RNAi reporters, conditional knockdown, and cell fate controls (Figure 8.4). For example, DNA-mediated siRNA switches consist of DNA–RNA hybrids that may be suitable for the systemic delivery of siRNA. *In vivo* (mouse) studies have demonstrated that these switches promote degradation resistance in the

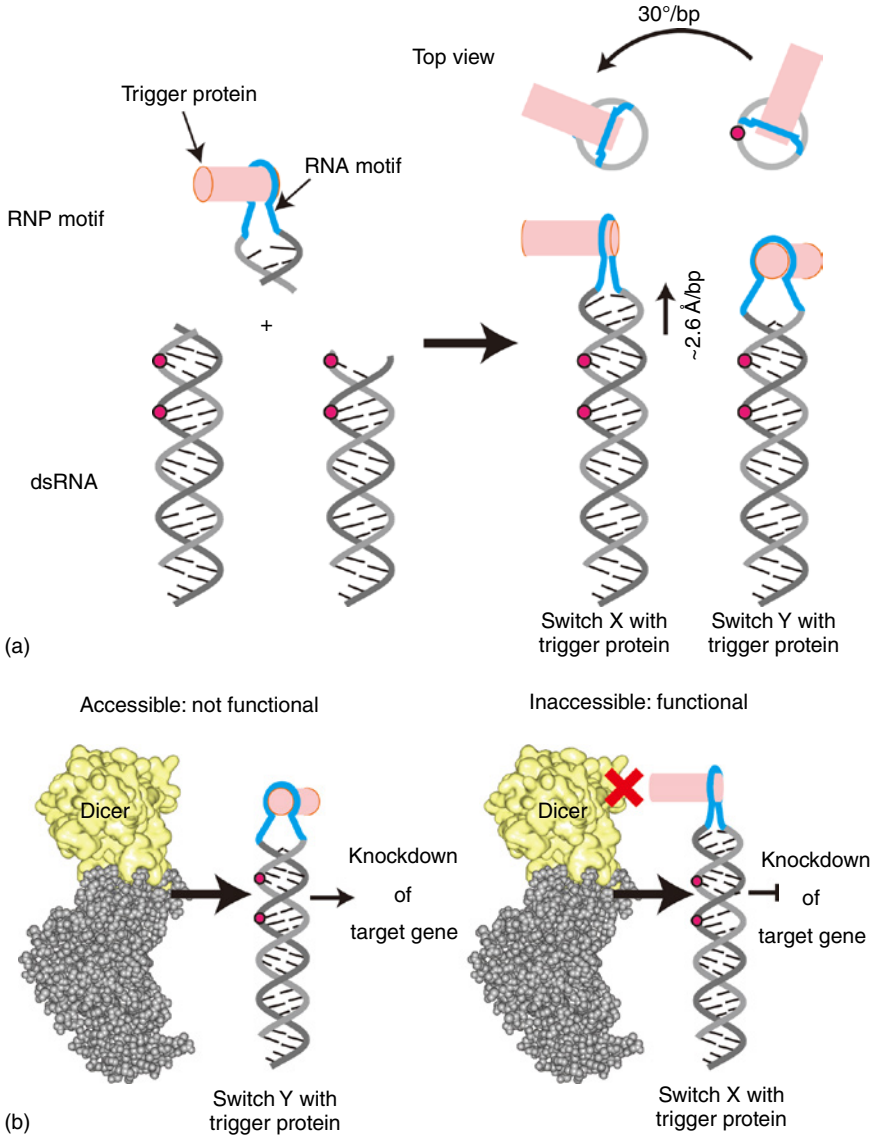


Figure 8.3 3D design scheme of the protein-responsive shRNA switch. (a) Protein-responsive shRNA switches are three-dimensionally designed by connecting an RNP motif to the corresponding dsRNA *in silico*. After specifying the position of the devices with reference to the Dicer cleavage sites and comparing the locations of the bound proteins of the X and Y switches, the bound trigger protein on switch X rotates ~30° in a counterclockwise direction around the axis of the dsRNA and is located ~2.6 Å more distant from the Dicer cleavage sites than switch Y. (b) Dicer can access and process switch Y via trigger protein binding; the switch then induces the knockdown of its target gene via RNAi (left). Dicer is inaccessible to switch X in the presence of the trigger protein because the RNP interaction faces Dicer and inhibits its access (right). The prevention of Dicer function causes the derepression of gene knockdown.

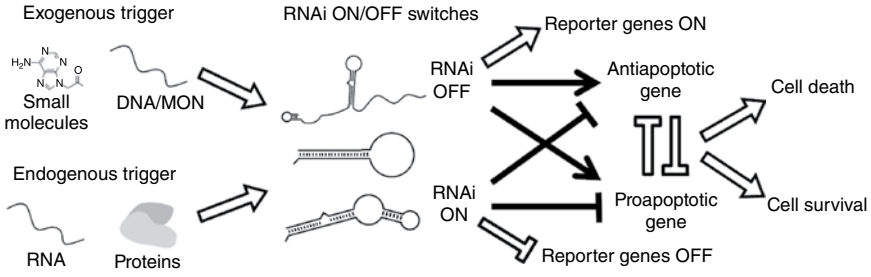


Figure 8.4 Applications of RNAi switches. RNAi ON/OFF switches can be applied to RNAi reporter and cell fate conversion systems that respond to specific trigger molecules.

bloodstream, efficient uptake by tumors, and reassociation-triggered activation of split siRNA functioning [18]. MON-triggered pre-miRNA switches can be applied to reporters of activated siRNA molecules in individual cells. Equal amounts of MON (including siRNA and pre-miRNA targeting for fluorescent proteins) are produced in the nucleus when Drosha cleaves MON-bound pre-miRNA switches [15]. The levels of RNAi and siRNA molecules can thus be more precisely monitored and visualized than with co-transfection of target and reference (fluorescent protein target) siRNA. Protein-triggered shRNA switches can be applied to control cell fate. Protein-triggered shRNA switches can respond to human U1A and NFκB protein expression within cells [26]. L7Ae-triggered shRNA switches were shown to control human cell fate by regulating the balance between proapoptotic (Bim) and antiapoptotic (Bcl-xL) protein molecules via the knockdown of antiapoptotic proteins (Bcl-xL) and by determining the status of mitochondrial-dependent apoptosis pathways. The expression of L7Ae determines cell survival.

8.5 Future Perspectives

Recent intensive research has resulted in the development of RNAi switches that are triggered by multiple chemicals and biomacromolecules. To improve the ability of RNAi switches to rewire gene regulatory networks, however, there are several challenges to overcome regarding switch efficiency and the variety of specific trigger molecules for other RNA switches. For example, protein-triggered shRNA switches require high plasmid expression levels of the trigger protein. To generate switches that respond to endogenous protein molecules, designed RNA devices must efficiently and selectively detect target proteins [28]. Extra signal amplification systems such as synthetic positive feedback loops may be required to generate sufficient protein signals. Additionally, an orthogonal RNA–protein-binding pair that does not interfere with natural RNA or protein molecules is desirable to sense target proteins without inducing side effects. Thus, it is important to develop an automated and easy selection method to generate such specific RNA–protein-binding pairs from RNA motif libraries.

Definitions

RNAi RNA interference. The term is used synonymously with knockdown
RNAi switch RNAi-inducible short RNA component, the function of which is modulated by trigger molecules

References

1 Fellmann, C. and Lowe, S.W. (2014) Stable RNA interference rules for silencing. *Nat. Cell Biol.*, **16**, 10–18.
2 Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A. *et al.* (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, **391**, 806–811.
3 Wianny, F. and Zernicka-Goetz, M. (2000) Specific interference with gene function by double-stranded RNA in early mouse development. *Nat. Cell Biol.*, **2**, 70–75.
4 Elbashir, S.M., Harborth, J., Lendeckel, W., Yalcin, A. *et al.* (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, **411**, 494–498.
5 Agrawal, N., Dasaradhi, P.V.N., Mohmmed, A., Malhotra, P. *et al.* (2003) RNA interference: biology, mechanism, and applications. *Microbiol. Mol. Biol. Rev.*, **67**, 657–685.
6 Koponen, J.K., Kankkonen, H., Kannasto, J., Wirth, T. *et al.* (2003) Doxycycline-regulated lentiviral vector system with a novel reverse transactivator rtTA2S-M2 shows a tight control of gene expression *in vitro* and *in vivo*. *Gene Ther.*, **10**, 459–466.
7 Urlinger, S., Baron, U., Thellmann, M., Hasan, M.T. *et al.* (2000) Exploring the sequence space for tetracycline-dependent transcriptional activators: novel mutations yield expanded range and sensitivity. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 7963–7968.
8 Gossen, M. and Bujard, H. (1992) Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 5547–5551.
9 Deans, T.L., Cantor, C.R., and Collins, J.J. (2007) A tunable genetic switch based on RNAi and repressor proteins for regulating gene expression in mammalian cells. *Cell*, **130**, 363–372.
10 An, C.-I., Trinh, V.B., and Yokobayashi, Y. (2006) Artificial control of gene expression in mammalian cells by modulating RNA interference through aptamer–small molecule interaction. *RNA*, **12**, 710–716.
11 Beisel, C.L., Bayer, T.S., Hoff, K.G., and Smolke, C.D. (2008) Model-guided design of ligand-regulated RNAi for programmable control of gene expression. *Mol. Syst. Biol.*, **4**, 224.
12 Kumar, D., An, C.-I., and Yokobayashi, Y. (2009) Conditional RNA interference mediated by allosteric ribozyme. *J. Am. Chem. Soc.*, **131**, 13906–13907.
13 Masu, H., Narita, A., Tokunaga, T., Ohashi, M. *et al.* (2009) An activatable siRNA probe: trigger-RNA-dependent activation of RNAi function. *Angew. Chem. Int. Ed.*, **48**, 9481–9483.

14 Beisel, C.L., Chen, Y.Y., Culler, S.J., Hoff, K.G. *et al.* (2011) Design of small molecule-responsive microRNAs based on structural requirements for Drosha processing. *Nucleic Acids Res.*, **39**, 2981–2994.

15 Kumar, D., Kim, S.H., and Yokobayashi, Y. (2011) Combinatorially inducible RNA interference triggered by chemically modified oligonucleotides. *J. Am. Chem. Soc.*, **133**, 2783–2788.

16 Saito, H., Fujita, Y., Kashida, S., Hayashi, K. *et al.* (2011) Synthetic human cell fate regulation by protein-driven RNA switches. *Nat. Commun.*, **2**, 160.

17 Kashida, S., Inoue, T., and Saito, H. (2012) Three-dimensionally designed protein-responsive RNA devices for cell signaling regulation. *Nucleic Acids Res.*, **40**, 9369–9378.

18 Afonin, K.A., Viard, M., Martins, A.N., Lockett, S.J. *et al.* (2013) Activation of different split functionalities on re-association of RNA-DNA hybrids. *Nat. Nanotechnol.*, **8**, 296–304.

19 Velagapudi, S.P., Gallo, S.M., and Disney, M.D. (2014) Sequence-based design of bioactive small molecules that target precursor microRNAs. *Nat. Chem. Biol.*, **10**, 291–297.

20 Zimmermann, G.R., Wick, C.L., Shields, T.P., Jenison, R.D. *et al.* (2000) Molecular interactions and metal binding in the theophylline-binding core of an RNA aptamer. *RNA*, **6**, 659–667.

21 Zhang, H., Kolb, F.A., Brondani, V., Billy, E. *et al.* (2002) Human Dicer preferentially cleaves dsRNAs at their termini without a requirement for ATP. *EMBO J.*, **21**, 5875–5885.

22 Beisel, C.L. and Smolke, C.D. (2009) Design principles for riboswitch function. *PLoS Comput. Biol.*, **5**, e1000363.

23 Müller, M., Weigand, J.E., Weichenrieder, O., and Suess, B. (2006) Thermodynamic characterization of an engineered tetracycline-binding riboswitch. *Nucleic Acids Res.*, **34**, 2607–2617.

24 Ogawa, A. (2011) Rational design of artificial riboswitches based on ligand-dependent modulation of internal ribosome entry in wheat germ extract and their applications as label-free biosensors. *RNA*, **17**, 478–488.

25 Salis, H.M., Mirsky, E.A., and Voigt, C.A. (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.*, **27**, 946–950.

26 Kashida, S. and Saito, H. (2014) A three-dimensional design strategy for a protein-responsive shRNA switch. *Methods Mol. Biol.*, **1111**, 269–286.

27 Gu, S., Jin, L., Zhang, Y., Huang, Y. *et al.* (2012) The loop position of shRNAs and pre-miRNAs is critical for the accuracy of dicer processing in vivo. *Cell*, **151**, 900–911.

28 Kawasaki, S., Fujita, Y., Nagaike, T., Tomita, K. *et al.* (2017) Synthetic mRNA devices that detect endogenous proteins and distinguish mammalian cells. *Nucleic Acids Res.*, 1–13.

9

Small Molecule-Responsive RNA Switches (Bacteria): Important Element of Programming Gene Expression in Response to Environmental Signals in Bacteria

Yohei Yokobayashi

Okinawa Institute of Science and Technology Graduate University, Nucleic Acid Chemistry and Engineering Unit
Onna-son, Okinawa 9040415, Japan

9.1 Introduction

Engineering small molecule-responsive RNA switches in bacteria was motivated by the discovery of natural prokaryotic riboswitches in 2002 [1–3]. These endogenous RNA *cis*-regulatory elements are usually found in the 5′ untranslated region (UTR) of prokaryotic mRNAs and modulate gene expression in response to various metabolites [4]. A typical riboswitch contains an aptamer domain that is responsible for metabolite binding and an expression platform that facilitates a ligand-dependent structural change that influences gene expression. For example, a metabolite-mediated structural change may alter the accessibility of the ribosome binding site (RBS), which results in a change in translation efficiency, or dictate the formation of a transcription terminator structure (a stem-loop followed by a short poly(U) tract) that results in premature termination of the transcript. The ability of these riboswitches to control gene expression in response to small molecules of biological or synthetic origin can be very useful in synthetic biology and metabolic engineering. This section provides a brief overview of the previous major efforts to engineer small molecule-responsive RNA switches in bacteria.

9.2 Design Strategies

9.2.1 Aptamers

An RNA aptamer that specifically binds to a desired small molecule ligand is a prerequisite to engineering riboswitches. Most published synthetic riboswitches have used known aptamers selected *in vitro* or metabolite-binding aptamers found in natural riboswitches. At least one group has performed *in vitro* selection to develop novel aptamers specifically for riboswitch applications in bacteria [5]. While researchers have successfully performed *in vitro* selection to discover

Synthetic Biology: Parts, Devices and Applications, First Edition. Edited by Christina Smolke.

aptamers against numerous small molecules [6], isolating new aptamers for novel targets for synthetic riboswitch applications is still likely to remain a challenge in itself [7]. Although *in vitro* selection often yields aptamers with respectable affinity and specificity, a major challenge when using them in the cellular context is the difficulty in predicting how the affinity, stability, or folding of the aptamers is altered inside the cells. To mitigate such uncertainty, Gallivan and coworkers used a pool of affinity-enriched aptamers from an *in vitro* selection, rather than few isolated aptamer clones, in their effort to engineer riboswitches that respond to an herbicide in *Escherichia coli* [5].

In another notable effort, Dixon and coworkers conducted an *in vivo* screen to modify a natural aptamer to recognize an alternative synthetic analog [8]. Although this strategy is likely to be limited for engineering aptamers for a ligand that are structurally similar to an existing ligand, it represents a viable alternative route to obtain a set of orthogonal riboswitches.

9.2.2 Screening and Genetic Selection

With an aptamer in hand, designing an appropriate expression platform becomes the primary challenge in engineering small molecule-responsive riboswitches. By far, the most successful strategies have employed some form of medium- to high-throughput screening or genetic selection at this stage to discover functional riboswitches with desired characteristics. Generally, a short stretch of sequence near an aptamer embedded in the 5′ UTR is randomized with an anticipation that a subset of those sequences will function as an expression platform. This pool of riboswitch mutants is subjected to suitable screening or selection steps to enrich functional riboswitches and to eventually isolate individual clones.

Genetic selection enables rapid enrichment of potential riboswitches from a large population ($>10^5$) of mutants by coupling the survival or growth of the bacteria with those expressing functional riboswitch mutants. Nomura and Yokobayashi isolated riboswitches entirely through genetic selection for the first time [9]. This was achieved by the use of tetracycline antiporter (*tetA*) as a selection marker to enable both ON and OFF selection. In this system, ON cells are selected using tetracycline and OFF cells are selected using $NiCl_2$ added to the culture media [10]. The group later improved the method by adding a fluorescent reporter gene (GFPuv) as a translational fusion to TetA to enable rapid screening of the genetically selected mutants [11].

Alternatively, Topp and Gallivan devised a selection strategy based on cell motility by coupling the riboswitch output with the expression of *cheZ*, which confers cell motility when expressed in a Δ*cheZ* host [12]. In this method, cells are physically isolated on a semisoft agar plate based on their motility.

Although genetic selection enables examination of relatively large number of mutants primarily limited by the transformation efficiency, it is often difficult to fine-tune the selection pressures to engineer devices with precise characteristics. A complementary strategy is to employ a reporter gene such as green fluorescent protein (GFP) and quantitatively measure the riboswitch performance

of individual mutants. The Gallivan and the Hartig groups, among others, have taken this approach by evaluating hundreds to thousands of riboswitch clones by reporter gene assay [13–15]. Although more laborious and costly compared with genetic selection, screening of individual clones provides quantitative characteristics of every mutant evaluated (ON and OFF expression levels). More recently, fluorescence-activated cell sorting (FACS) has been used to further increase throughput [16, 17].

9.2.3 Rational Design

Despite the extensive research on the natural riboswitch mechanisms and structures, rational or computational design of synthetic bacterial riboswitches has been few and far between. An earlier example by Suess *et al*. highlighted the potential of rationally engineering ligand-induced structural shift to regulate bacterial gene expression [18]. More recently, computationally driven designs of bacterial RNA switches based on ribozymes [19] and transcriptional regulation [20] have emerged. However, some level of experimental feedback is still expected to be essential due to the complexity of parameters that influence the performance of these RNA devices in living cells.

9.3 Mechanisms

9.3.1 Translational Regulation

Translational regulation by bacterial riboswitches involves a change in the local structure of the ribosome binding site (RBS) upon ligand binding. RBS within a stable structure generally hinders ribosome access and results in repressed translation. Engineering such ligand-induced structural changes, however, is not trivial, and screening or selection is often used in the process. In some cases, naive randomization of the nucleotides peripheral to the RBS followed by screening or selection was sufficient for isolation of suitable expression platforms [9, 13]. In other cases, riboswitch libraries were carefully designed to predispose the riboswitch mutants to undergo a specific structural shift. An example of the latter is shown in Figure 9.1a where the RBS was strategically placed to form a putative stem at the base of the aptamer upon ligand binding so that the riboswitch negatively responds to the aptamer ligand [21].

In another strategy, the RBS was placed so that it becomes accessible only when the hammerhead ribozyme self-cleaves, and the aptamer was inserted in one of the stem-loops of the ribozyme to control its activity (Figure 9.1b) [14, 15, 22]. In this strategy, because the translation efficiency is directly coupled to the ribozyme activity, small molecule response is actually engineered at the level of the aptamer–ribozyme hybrid, or aptazyme. The Hartig group has exploited the aptazyme strategy further by adapting them to control other translational components such as tRNA [23] and rRNA [24] to construct small molecule-responsive RNA switches in *E. coli*.

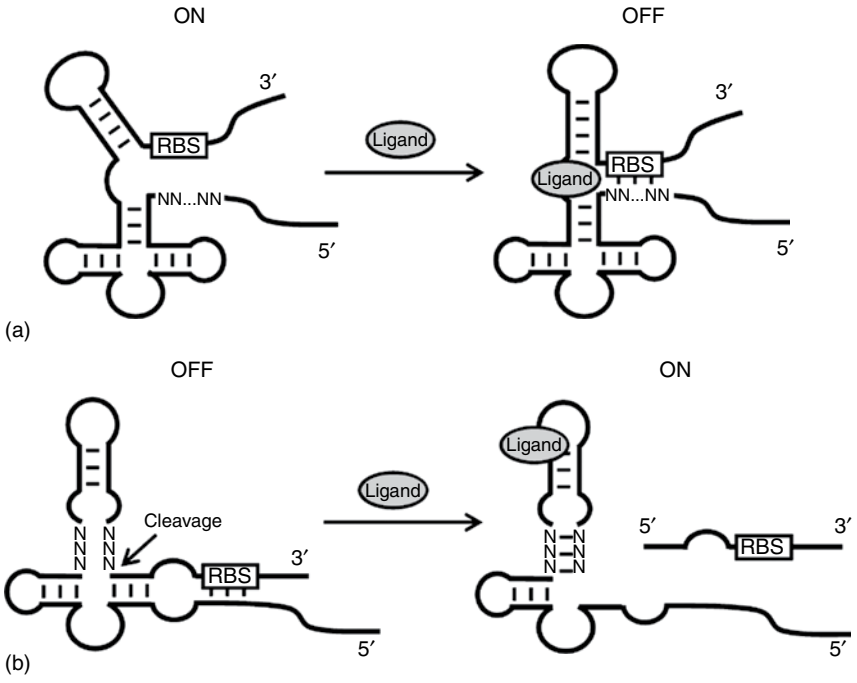


Figure 9.1 (a,b) Examples of synthetic riboswitch libraries.

9.3.2 Transcriptional Regulation

Relative to translationally regulated riboswitches, transcriptional regulation has not been extensively exploited in engineered riboswitches. Although this could be partly due to the complex mechanisms observed in the transcriptionally regulated natural riboswitches that involve folding kinetics of multiple RNA elements [25], several recent publications indicate that it is possible to engineer such riboswitches. Wachsmuth *et al*. used computational tools to rationally design a transcriptionally regulated riboswitch using a theophylline aptamer [20]. After some iterative improvements, they obtained a riboswitch with a respectable ON/OFF ratio of 6.5 in *E. coli*.

Alternatively, Qi and coworkers engineered *trans*-acting small noncoding RNAs (ncRNA) that function by transcriptionally regulating the target gene [26]. Their system is based on the antisense RNA-mediated transcription attenuation observed in the staphylococcal plasmid pT181 [27, 28]. By strategically fusing an aptamer and the ncRNA in tandem and screening of mutants in *E. coli*, the group successfully isolated small molecule-regulated *trans*-acting RNA switches.

More recently, Ceres and coworkers discovered that certain expression platforms of transcriptionally regulated riboswitches can accommodate different natural and synthetic aptamers without losing the gene regulatory function [29]. They were also able to qualitatively tune the device characteristics by adjusting the strength of the key stem sequences in a predictable fashion.

9.4 Complex Riboswitches

Although the majority of natural riboswitches provides a simple yet efficient means of metabolite-controlled gene expression, few noncanonical riboswitches that contain two aptamers have been reported [30–32]. These riboswitches have been shown or predicted to exhibit functions more complex compared with the single aptamer riboswitches, such as cooperative response to a ligand [30] or Boolean logic response to two distinct metabolites [31].

Several synthetic mimics of these complex bacterial riboswitches have been constructed. Sharma *et al.* constructed riboswitches that function as AND and NAND logic gates in response to theophylline and thiamine pyrophosphate (TPP) by *tetA* genetic selection [33]. More recently, Muranaka and Yokobayashi combined two independently optimized TPP riboswitches into the same 5′ UTR to construct a "band-pass" riboswitch that activates gene expression within a limited range of the ligand concentration [34].

In an alternative approach, Klauser *et al.* recently combined multiple ribozyme-based switches to create logic gates [35]. Qi *et al.* co-expressed two allosteric *trans*-acting ncRNA regulators to demonstrate a NOR logic gate with a small molecule and a protein as inputs [26].

9.5 Conclusions

As briefly summarized previously, synthetic riboswitches are attractive tools for interfacing bacterial synthetic circuits with small molecules of synthetic or natural origins. In particular, the demonstrated versatility of RNA aptamers to recognize a wide variety of molecules is of practical importance although adapting *in vitro* selected aptamers to intracellular applications remains a technical challenge. It is also noteworthy that complex functions that include molecular recognition, gene regulation, and, in some cases, multiple signal integration can all be encoded within one or few short segments of RNA, whereas equivalent switches and circuits based on protein transcription factors would require much larger genetic information.

Keywords with Definitions

Untranslated region (UTR) The sequences within an mRNA that do not code for a protein

Aptamer A nucleotide sequence (e.g., RNA) that is capable of binding specific target molecules such as small molecules and proteins

Expression platform A sequence within a riboswitch that is responsible for ligand-mediated structural change resulting in gene regulation

Riboswitch A stretch of RNA sequence mostly found in the 5′ UTR of bacterial mRNAs that binds a metabolite through an aptamer and regulates expression of the *cis*-gene

Ribozyme An RNA sequence capable of catalyzing a chemical reaction such as self-cleavage

References

1 Nahvi, A., Sudarsan, N., Ebert, M.S., Zou, X., Brown, K.L. *et al.* (2002) Genetic control by a metabolite binding mRNA. *Chem. Biol.*, **9**, 1043.

2 Winkler, W., Nahvi, A., and Breaker, R.R. (2002) Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature*, **419**, 952–956.

3 Mironov, A.S., Gusarov, I., Rafikov, R., Lopez, L.E., Shatalin, K. *et al.* (2002) Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell*, **111**, 747–756.

4 Serganov, A. and Nudler, E. (2013) A decade of riboswitches. *Cell*, **152**, 17–24.

5 Sinha, J., Reyes, S.J., and Gallivan, J.P. (2010) Reprogramming bacteria to seek and destroy an herbicide. *Nat. Chem. Biol.*, **6**, 464–470.

6 Stoltenburg, R., Reinemann, C., and Strehlitz, B. (2007) SELEX – a (r) evolutionary method to generate high-affinity nucleic acid ligands. *Biomol. Eng*, **24**, 381–403.

7 Carothers, J.M., Goler, J.A., Kapoor, Y., Lara, L., and Keasling, J.D. (2010) Selecting RNA aptamers for synthetic biology: investigating magnesium dependence and predicting binding affinity. *Nucleic Acids Res.*, **38**, 2736–2747.

8 Dixon, N., Duncan, J.N., Geerlings, T., Dunstan, M.S., McCarthy, J.E. *et al.* (2010) Reengineering orthogonally selective riboswitches. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 2830–2835.

9 Nomura, Y. and Yokobayashi, Y. (2007) Reengineering a natural riboswitch by dual genetic selection. *J. Am. Chem. Soc.*, **129**, 13814–13815.

10 Podolsky, T., Fong, S.T., and Lee, B.T. (1996) Direct selection of tetracycline-sensitive *Escherichia coli* cells using nickel salts. *Plasmid*, **36**, 112–115.

11 Muranaka, N., Sharma, V., Nomura, Y., and Yokobayashi, Y. (2009) An efficient platform for genetic selection and screening of gene switches in *Escherichia coli*. *Nucleic Acids Res.*, **37**, e39.

12 Topp, S. and Gallivan, J.P. (2008) Random walks to synthetic riboswitches – a high-throughput selection based on cell motility. *ChemBioChem*, **9**, 210–213.

13 Lynch, S.A., Desai, S.K., Sajja, H.K., and Gallivan, J.P. (2007) A high-throughput screen for synthetic riboswitches reveals mechanistic insights into their function. *Chem. Biol.*, **14**, 173–184.

14 Wieland, M., Benz, A., Klauser, B., and Hartig, J.S. (2009) Artificial ribozyme switches containing natural riboswitch aptamer domains. *Angew. Chem. Int. Ed.*, **48**, 2715–2718.

15 Wieland, M. and Hartig, J.S. (2008) Improved aptazyme design and in vivo screening enable riboswitching in bacteria. *Angew. Chem. Int. Ed.*, **47**, 2604–2607.

16 Fowler, C.C., Brown, E.D., and Li, Y. (2008) A FACS-based approach to engineering artificial riboswitches. *ChemBioChem*, **9**, 1906–1911.

17 Lynch, S.A. and Gallivan, J.P. (2009) A flow cytometry-based screen for synthetic riboswitches. *Nucleic Acids Res.*, **37**, 184–192.

18 Suess, B., Fink, B., Berens, C., Stentz, R., and Hillen, W. (2004) A theophylline responsive riboswitch based on helix slipping controls gene expression in vivo. *Nucleic Acids Res.*, **32**, 1610–1614.

19 Carothers, J.M., Goler, J.A., Juminaga, D., and Keasling, J.D. (2011) Model-driven engineering of RNA devices to quantitatively program gene expression. *Science*, **334**, 1716–1719.

20 Wachsmuth, M., Findeiss, S., Weissheimer, N., Stadler, P.F., and Morl, M. (2013) De novo design of a synthetic riboswitch that regulates transcription termination. *Nucleic Acids Res.*, **41**, 2541–2551.

21 Muranaka, N., Abe, K., and Yokobayashi, Y. (2009) Mechanism-guided library design and dual genetic selection of synthetic OFF riboswitches. *ChemBioChem*, **10**, 2375–2381.

22 Ogawa, A. and Maeda, M. (2008) An artificial aptazyme-based riboswitch and its cascading system in *E. coli*. *ChemBioChem*, **9**, 206–209.

23 Berschneider, B., Wieland, M., Rubini, M., and Hartig, J.S. (2009) Small-molecule-dependent regulation of transfer RNA in bacteria. *Angew. Chem. Int. Ed.*, **48**, 7564–7567.

24 Wieland, M., Berschneider, B., Erlacher, M.D., and Hartig, J.S. (2010) Aptazyme-mediated regulation of 16S ribosomal RNA. *Chem. Biol.*, **17**, 236–242.

25 Haller, A., Souliere, M.F., and Micura, R. (2011) The dynamic nature of RNA as key to understanding riboswitch mechanisms. *Acc. Chem. Res.*, **44**, 1339–1348.

26 Qi, L., Haurwitz, R.E., Shao, W., Doudna, J.A., and Arkin, A.P. (2012) RNA processing enables predictable programming of gene expression. *Nat. Biotechnol.*, **30**, 1002–1006.

27 Brantl, S. and Wagner, E.G. (2000) Antisense RNA-mediated transcriptional attenuation: an in vitro study of plasmid pT181. *Mol. Microbiol.*, **35**, 1469–1482.

28 Novick, R.P., Iordanescu, S., Projan, S.J., Kornblum, J., and Edelman, I. (1989) pT181 plasmid replication is regulated by a countertranscript-driven transcriptional attenuator. *Cell*, **59**, 395–404.

29 Ceres, P., Garst, A.D., Marcano-Velázquez, J.G., and Batey, R.T. (2013) Modularity of select riboswitch expression platforms enables facile engineering of novel genetic regulatory devices. *ACS Synth. Biol.* doi: 10.1021/sb4000096

30 Mandal, M., Lee, M., Barrick, J.E., Weinberg, Z., Emilsson, G.M. *et al.* (2004) A glycine-dependent riboswitch that uses cooperative binding to control gene expression. *Science*, **306**, 275–279.

31 Sudarsan, N., Hammond, M.C., Block, K.F., Welz, R., Barrick, J.E. *et al.* (2006) Tandem riboswitch architectures exhibit complex gene control functions. *Science*, **314**, 300–304.

32 Welz, R. and Breaker, R.R. (2007) Ligand binding and gene control characteristics of tandem riboswitches in *Bacillus anthracis*. *RNA*, **13**, 573–582.

33 Sharma, V., Nomura, Y., and Yokobayashi, Y. (2008) Engineering complex riboswitch regulation by dual genetic selection. *J. Am. Chem. Soc.*, **130**, 16310–16315.

34 Muranaka, N. and Yokobayashi, Y. (2010) A synthetic riboswitch with chemical band-pass response. *Chem. Commun. (Camb)*, **46**, 6825–6827.

35 Klauser, B., Saragliadis, A., Auslander, S., Wieland, M., Berthold, M.R. *et al.* (2012) Post-transcriptional Boolean computation by combining aptazymes controlling mRNA translation initiation and tRNA activation. *Mol. Biosyst.*, **8**, 2242–2248.

10

Programming Gene Expression by Engineering Transcript Stability Control and Processing in Bacteria

Jason T. Stevens and James M. Carothers

University of Washington, Center for Synthetic Biology, Molecular Engineering and Sciences Institute, Departments of Chemical Engineering and Bioengineering, 4000 15th Ave NE, Seattle, WA 98195-1654, USA

Through control of messenger RNA stability, bacteria are able to process information, respond to changing conditions, and maintain homeostasis. Many of the naturally occurring mechanisms for transcript stability control (TSC) have been elucidated, and a number of studies have leveraged this understanding to demonstrate that transcript stability can be engineered to control static and dynamic gene expression. Collectively, that body of work represents a foundation for developing new forward-engineering approaches that harness mechanistic understanding to build predictive computational models to guide the development of large-scale genetic devices based on TSC and other means. Further increasing our understanding of RNA degradation pathways and mechanisms will also improve the ability to anticipate how undesired variations in transcript stability may confound device output goals and frustrate engineering efforts. Here, we discuss the current state of the art and identify routes for using TSC to design increasingly large and complex synthetic biological systems.

10.1 An Introduction to Transcript Control

10.1.1 Why Consider Transcript Control?

In naturally occurring biological systems, RNA-based genetic control mechanisms play crucial roles in regulating cellular functions. Genome-wide studies of bacterial transcript half-lives [1, 2] have underscored the importance that control over transcript stability plays in enabling bacteria to process information, respond to changing cellular and environmental conditions, and, ultimately, maintain homeostasis. Bacterial transcripts are known to persist for times that vary in scale over orders of magnitude, from only a few seconds to an hour, or more. Nature uses several mechanisms to control transcript persistence [3–6], and experimental evidence has shown that these mechanisms can be engineered [7–10], providing a route to programming static and dynamic gene expression [11]. Developing technologies for designing variations in transcript stability

Synthetic Biology: Parts, Devices and Applications, First Edition. Edited by Christina Smolke.

could, therefore, increase the speed with which synthetic biological systems can be created for applications in basic science; for the production of renewable chemicals, fuels, and materials for global health; and for the development of new therapeutic agents. Even when transcript stability is not an explicit aspect of a given genetic control device (e.g., ribosome binding site (RBS) control of translation initiation), unknown or poorly characterized effects on transcript degradation may affect genetic device outputs. It is therefore important to regard transcript stability through two lenses: as a "tuning knob" for predictably controlling gene expression dynamics and as a confounding factor if unaccounted for in genetic device design.

In this chapter, we describe current understanding of transcript stability and processing for designing and engineering genetic expression devices with predictable functions. In Section 10.1, we consider the machinery that controls transcript stability within bacteria, with specific focus on *Escherichia coli*. Section 10.2 examines efforts to utilize this machinery for controlling gene expression dynamics. In Section 10.3, we consider ways of managing transcript stability to reduce unintentional and confounding effects. Section 10.4 details possible strategies for controlling transcript stability and points to future research directions in computation and wet-lab experimentation that may lead to design technologies for rapidly engineering genetic devices. The final section, Section 10.5, will provide a summary of the chapter.

10.1.2 The RNA Degradation Process in *E. coli*

RNA is degraded through multistep pathways that can begin as soon as a transcript has been synthesized by an RNA polymerase. Degradation of mRNA typically begins with a rate-limiting, RNase E-mediated phosphodiester bond cleavage event. RNase E cleavage is followed by subsequent rounds of $3' \rightarrow 5'$ degradation (*E. coli* has no known $5' \rightarrow 3'$ exoribonuclease) [4], carried out in concert with the degradosome, a collection of four enzymes – RNase E, RhlB, PNPase (polynucleotide phosphorylase), and enolase [3, 12–14] – that localizes to the membrane [15, 16].

RNase E [17], an endoribonuclease and a rate-limiting cleavage enzyme, is thought to bind and process transcripts via two mechanisms (Figure 10.1), the first of which is 5′ entry at a monophosphorylated end. It was discovered to prefer substrates with unpaired 5′ ends *in vivo* [18], and early *in vitro* analysis of RNase E activity showed a manyfold reduction in cleavage rate when three different RNAs were 5′ triphosphorylated (5′-PPP) instead of 5′ monophosphorylated (5′-P) [19]. The structure of the RNase E domain was later shown to have a binding pocket that cannot accommodate substrates larger than a 5′-P [20], explaining the selectivity toward 5′-P- versus 5′-PPP-terminated transcripts and the increased half-life of 5′ hydroxyl (5′-OH)-terminated transcripts [21]. As transcripts are synthesized natively with 5′-PPP, it was hypothesized, and later shown, that 5′-P RNAs are created in cells through the removal of the gamma- and beta-phosphate from 5′-PPP RNAs [21]. This conversion, which creates the direct substrates for RNase E cleavage, was found to be catalyzed by RppH, an RNA pyrophosphohydrolase [22].

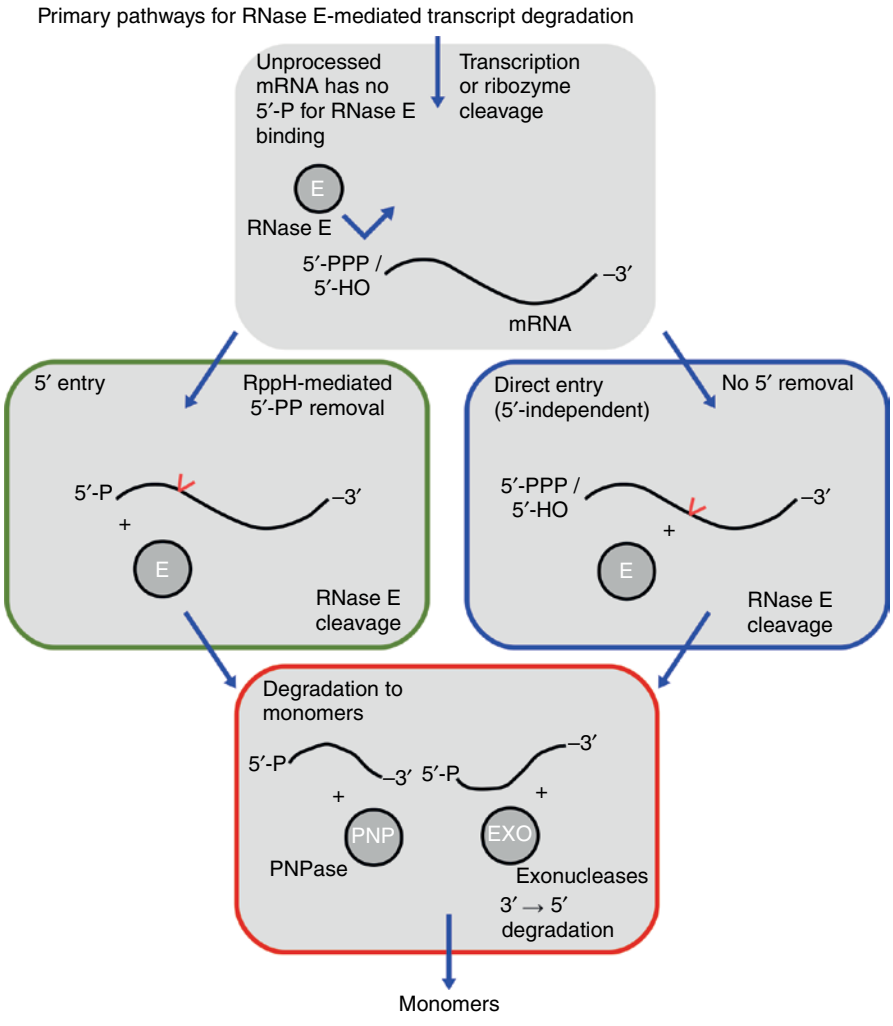


Figure 10.1 Primary routes for RNase E-mediated mRNA degradation. "5′ entry" is initiated when an mRNA undergoes 5′-PP removal, catalyzed by the pyrophosphohydrolase enzyme RppH, creating a 5′-P that can be recognized and bound by RNase E (shown by "+" symbol). "Direct entry" (at right) is 5′ independent entry by RNase E that occurs without recognition and binding to a 5′-P moiety. Following RNase E binding, an initial cleavage event generates 3′-OH- and 5′-P-terminated RNAs that are efficient substrates for 3′ → 5′ degradation to monomers and further rounds of RNase E binding and cleavage.

The second method of RNase E substrate recognition, often termed "direct entry," bypasses the 5′ end [23]. Experiment has demonstrated that mRNAs containing a putative 5′ hairpin to inhibit 5′ RNase E binding is still degraded in an RNase E-dependent manner [24], and the insertion of putative RNase E sites into the coding region decreased the stability of RNA with a 5′ hairpin [25]. Additionally, several RNAs have been identified that can be rapidly degraded by the RNase E catalytic domain even if they are not terminated with a 5′-P (i.e., if

they are 5′-PPP- or 5′-OH-terminated RNAs), thought to occur when RNase E binds directly to unpaired regions in the coding sequence [26].

Once RNase E is associated with the RNA, either through 5′-P binding or direct entry, it can scan the transcript and catalyze cleavage at the initial target site [27, 28], setting in motion the recruitment of the other members of the degradosome and further degradation by $3' \rightarrow 5'$ exoribonucleases and subsequent rounds of RNase E activity [29]. PNPase, a $3' \rightarrow 5'$ exoribonuclease, binds to polyadenylated 3′ mRNA ends [30]. RhlB is an adenosine triphosphate (ATP)-dependent helicase implicated in preparing RNA for RNase E and PNPase cleavage by removing secondary structure. When inhibited, RhlB no longer enhanced PNPase-mediated degradation in an ATP-dependent way [31], and when RhlB was deleted, *lacZ* mRNA was stabilized in a ribosome-free context by impaired RNase E cleavage at the 5′ end [32]. Enolase is the least well-understood member of the degradosome and is thought to have a role in metabolism-related transcript degradation [33].

Additional means of initiating degradation occur through RNase III cleavage and RNase G cleavage. RNase III is thought to primarily bind and cleave secondary structures [34], often in the context of rRNA maturation and decay [35]. RNase G, an RNase E homolog, is usually involved in 9S rRNA maturation but in a small number of cases initiates mRNA decay as well [36].

While this is not an exhaustive account of the mechanisms related to RNA decay in *E. coli*, the aforementioned mechanisms are responsible for the majority of messenger RNA decay [3] and are the most salient for programming variations in gene expression levels.

10.1.3 The Effects of Translation on Transcript Stability

The development of a complete mechanistic understanding of RNA degradation has been complicated by the effects that ribosomes and translation have on transcript stability (Figure 10.2a) [37]. For instance, ribosome binding has been found to attenuate RNase E cleavage in several studies. Incubation of the *ompA* mRNA with increasing molar excesses of 30S ribosomal subunits substantially reduced RNase E cleavage in the 5′ untranslated region (UTR) [38]. *lacZ* transcript half-life was correlated with β-galactosidase enzyme activity (a proxy for translation efficiency) when changes were made to the RBS [39], suggesting that ribosome occupancy positively influences transcript half-life. Taken together, these results support a simple steric hindrance model where the presence of ribosomes on an mRNA inhibits RNase E binding and cleavage [37].

Moreover, because translation is co-transcriptional in bacteria, the transcription rate can influence susceptibility to RNase E cleavage by determining the length of exposed transcript. If transcription outpaces the rate of ribosome binding and translation initiation, much of the transcript, including potential RNase E binding sites, will be exposed. A study with the *lacZ* gene and mutant T7 bacteriophage polymerases in *E. coli* showed an inverse correlation between β-galactosidase activity and the rate of T7 RNA transcription, a trend that was RNase E dependent [40]. Experiments using premature stop codons to render

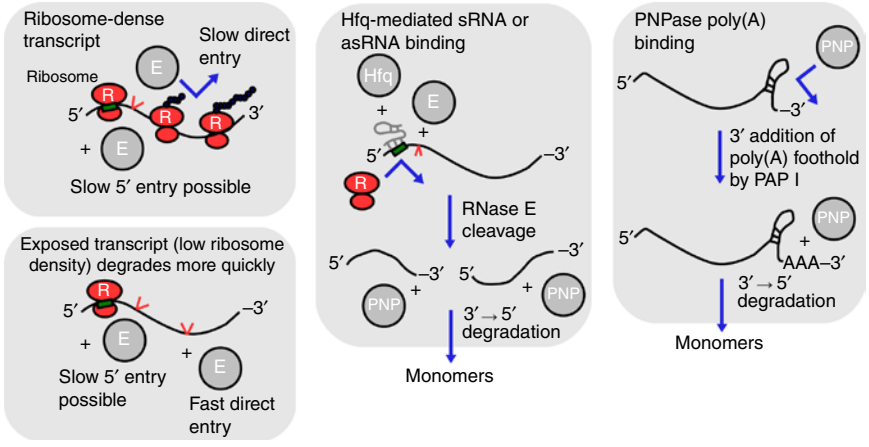


Figure 10.2 Naturally occurring transcript stability control mechanisms. (a) mRNAs that are highly occupied by translating ribosomes RNA will have occluded sites for RNase E 5′ entry and direct entry, leading to relatively long transcript half-life and high levels of gene expression. Exposed transcripts (i.e., with lower ribosome density), such as mRNAs transcribed with bacteriophage polymerases with fast elongation rates, are more susceptible to RNase E attack due to a lack of occluding ribosomes. (b) sRNA and asRNA operate through Hfq-mediated binding to the RBS (green box) and/or start codon region of a target mRNA, which prevents ribosome docking and likely recruits RNase E to the transcript. (c) The addition of poly(A) tails to a transcript, usually by poly(A) polymerase (PAP I), creates a foothold for binding by polynucleotide phosphorylase (PNPase), a 3′ → 5′ exoribonuclease.

transcripts ribosome-less at known RNase E cleavage sites showed decreased transcript stability [25, 41].

The complex interplay of other transcript-related mechanisms that affect transcript degradation is even less well understood. Ribosomal pausing can lead to cleavage by unknown ribonuclease activity. The subcellular localization of individual transcripts also affects ribosomal occupancy, which in turn affects RNA degradation [37].

10.1.4 Structural and Noncoding RNA-Mediated Transcript Control

In most cases, RNase E must bind the transcript – either at the 5′ end or at a single-stranded interior region – to initiate cleavage and begin degradation. This implies that anything affecting RNase E binding will also affect transcript stability. Ribosomes, as explained before, are one such factor. RNA secondary structures, or other stable base pairings that prevent access to the 5′ end or internal RNase E sites, are therefore expected to reduce RNase E binding and increase transcript stability (Figures 10.2b and 10.3a).

Antisense oligos that base-pair near the 5′ end of a transcript were shown to lower the rate of RNase E cleavage [19], and studies of naturally long-lived mRNA in *E. coli* have pointed to 5′ hairpins as a means of precluding RNase E docking and conferring transcript stability [31] [32]. A naturally occurring riboswitch, or

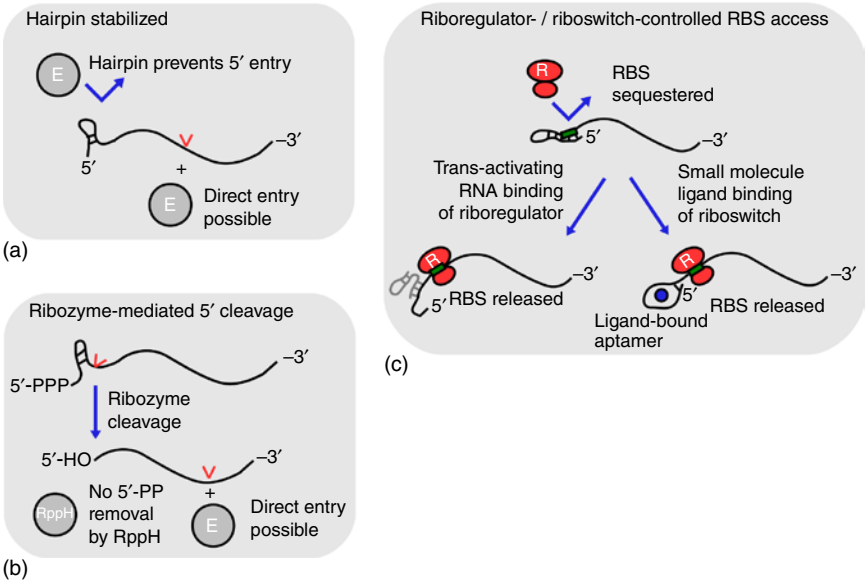


Figure 10.3 Examples of engineered transcript stability control. (a) Synthetic secondary structure hairpins within the 5′ UTR can increase half-life by preventing RNase E 5′ entry; direct entry by RNase E is still possible. (b) 5′ ribozyme-mediated transcript cleavage creates a 5′-OH not recognized by RppH and therefore cannot become a 5′-P for RNase E to bind; degradation of these processed transcripts occurs through RNase E direct entry. (c) Riboregulators and riboswitches typically work by *cis*-RNA sequestration of the RBS, which can be relieved by either *trans*-RNA binding to the *cis*-RN or ligand binding to the *cis*-RNA. These binding events free the RBS from the *cis*-RNA and therefore allow translation.

functional RNA that changes conformation upon ligand binding to dynamically sequester or present an RBS [42], has been discovered that also uncovers RNase E cleavage sites when bound to a target metabolite, lysine. As a result, lysine binding reduces the rate of translation initiation from the RBS and decreases transcript half-life, further reducing protein expression [28]. More generally, riboswitches are thought to decrease transcript stability in the RBS-sequestering state by precluding ribosome binding [28] (also see Section 10.1.3).

Like the antisense oligos, small RNA (sRNA) can create base pairing in the 5′ UTR and thus limit binding by ribosomes or RNase E. sRNA, an abundant form of regulatory noncoding RNA (ncRNA) in bacteria [43], is typically tens to hundreds of bases long. They usually function to enhance or repress ribosome binding by base pairing, via a short (10–20-bp) seed region, with the target mRNA at the Shine–Dalgarno (SD) and/or start codon regions of an mRNA [6, 44, 45]. In many cases, sRNA target binding is mediated by the RNA chaperone activity of the Hfq protein [46, 47]. Hfq deletion studies point to the importance of Hfq in sRNA action [47–49] and stability, with increased sRNA degradation by PNPase in stationary-phase *Hfq*-strains [50]. Hfq has been found to coprecipitate with nearly half of the known sRNAs of *Salmonella* [51] and with a quarter of the then known sRNAs in *E. coli* [52]. Interactions between some sRNA and mRNA can occur in the absence of Hfq, however [48, 53].

sRNAs can enhance translation when binding to the mRNA eliminates *cis*-acting secondary structure involving the SD sequence. In turn, ribosomes can bind to the SD and initiate translation, protecting the mRNA from degradation [54–57]. sRNAs can repress translation when binding to the mRNA occludes ribosome binding [58, 59], rendering the transcript susceptible to endonuclease cleavage due to a lack of protective ribosomes. In the latter case, sRNA binding seems to recruit RNase E through interaction with Hfq and the 5′-P of the sRNA[48], hastening degradation of both the sRNA and targeted mRNA. It is interesting to note that sRNA-mediated RBS occlusion is sufficient for downregulation; thus in some cases, degradation serves only to make the downregulation irreversible [53, 60].

10.1.5 Polyadenylation and Transcript Stability

Unlike in eukaryotes, polyadenylation of bacterial mRNA is not associated with transcript maturation and increased stability [5], but instead has been associated with mRNA destabilization (Figure 10.2c). Interestingly, although generally implicated in the degradation of nonfunctional or mutated RNAs as part of quality control mechanisms [61], there are several examples where polyadenylation is employed to modulate gene expression [5, 62, 63]. The half-life of *rpoS* mRNA in *E. coli* decreased when polyadenylated in the absence of RNase E, where polyadenylation depended on *pcnB* [64], the gene coding for poly(A) polymerase [65]. Poly(A) tails are used as footholds for exoribonucleases, such as PNPase, that bind the poly(A) tails and perform $3' \rightarrow 5'$ degradation [30, 62]. The half-life of three mRNA different transcripts increased when *pcnB* was knocked out, coinciding with poly(A) tails shortened up to 90% [66].

10.2 Synthetic Control of Transcript Stability

10.2.1 Transcript Stability Control as a "Tuning Knob"

As outlined in Section 10.1, transcript stability is determined through the collective impact of a multitude of sequence and structural features. The 5′ terminus identity (i.e., 5′-PPP vs 5′-P vs 5′-OH) and the presence of stable secondary structures within the 5′ UTR affect 5′ end accessibility by RppH and RNase E. Active translation creates steric hindrance and ribosome occlusion that reduces internal accessibility by RNase E. Finally, 3′ end accessibility by PNPase varies according to 3′ UTR secondary structure, polyadenylation state, and the presence or absence of sRNAs that mediate degradation. Because RNAs can be transcribed and degraded within the space of only a few minutes, variations in transcript stability can have dramatic effects on RNA levels. This implies that gene expression can be controlled quickly and dynamically by modulating the sequence and structural features that directly affect transcript stability. In naturally occurring systems, swings in transcript abundance allow cells to respond to changing conditions and, for instance, reestablish perturbed homeostasis or respond to the buildup of intra- or extracellular toxins [2]. TSC thus presents a

powerful platform for meeting system design goals for applications, such as metabolic pathway engineering or biosensing, that require the ability to generate specific levels of static gene expression or dynamic genetic outputs that change as the function of a targeted molecule. Many of the naturally occurring mechanisms can be tuned to program static levels of gene expression [67], and dynamic control [11] is possible if these static mechanisms are regulated by the binding activities of functional RNA structures evolved with *in vitro* selection to bind specific metabolites (e.g., RNA aptamers, or RNA aptamer-regulated ribozymes, aptazymes) [7].

Several TSC mechanisms have been used over the past 15 years in synthetic genetic systems. A small number of TSC mechanisms, namely, 5′ and 3′ UTR hairpins [67–72], 5′ UTR cleavage [7], and antisense RNA (asRNA)/sRNA binding [10, 53, 73–76], have been used to explicitly control transcript stability. (The systems developed using these mechanisms are discussed in detail in the following subsections.) Others were not explicit attempts to alter transcript stability [9, 10, 77, 78]. Rather, by changing ribosome binding and UTR structure, there were likely changes in RNA degradation, even though altered stability was not the chief actuator of control. Nevertheless, this substantial body of work has significantly advanced knowledge of RNA engineering that will undoubtedly be important in creating novel genetic control systems based on tuning mRNA stability. Moreover, this work has helped identify RNA components that are most easily engineered and understood and has reinforced the many strengths of RNA-based technologies, namely, low host metabolic burden [75], inherent orthogonality [76], and the evolvability [79, 80] of new components. With this work and advancing knowledge of degradation processes, transcript stability is poised to become a powerful means of genetic control, either on its own or as part of a larger control scheme.

Moving forward with increasing understanding of RNA device design principles and mechanistic understanding of degradation processes, it should be possible to formulate model-driven frameworks based on TSC mechanisms. Casting biochemical, mechanistic understanding of transcript degradation in terms of measurable and tunable design variables will enable us to take advantage of computational techniques to increase the speed of design, predictability, and scale of synthetic biological systems [7].

10.2.2 Secondary Structure at the 5′ and 3′ Ends

The earliest attempts to engineer the stability of transcripts in bacteria involved hindering ribonuclease' entry by adding stable hairpin secondary structures to the 5′ end of transcripts (Figure 10.3a) [67–69, 81, 134] or to the 3′ end [70, 71], followed by hairpins at both termini [72]. When a hairpin from the T7 *gene10* leader sequence was added to the 5′ end of *lacZ* in *E. coli*, β-galactosidase activity increased threefold, but only when RNase E was present, suggesting that the hairpin increased transcript half-life by reducing RNase E binding and cleavage rates [69]. A similar experiment also saw a threefold improvement in half-life after a 5′ hairpin addition [67]. Carrier and Keasling built a small library of 5′ hairpins that conferred an order-of-magnitude range in half-life, from 2 min up

to almost 20 min [81]. Similarly, 3′ hairpin introduction has been shown to inhibit degradation [70] and increased the penicillinase (*penP*) transcript half-life threefold in both *E. coli* and *Bacillus subtilis* [71].

These results demonstrated that transcript secondary structures inhibiting RNase E and exoribonuclease binding are useful tools for varying mRNA half-life in a static manner. Despite these successes, however, it has been difficult to control transcript stability in a quantitatively predictable manner through secondary structure engineering [72, 82]. In principle, it should be possible to develop more explicit design rules if the relationships between secondary structure folding kinetics, stability, and RNase E binding occlusion can be further developed. Cambray *et al*. combined experiment with kinetic RNA folding simulation analysis of a large number of transcriptional terminators to derive heuristics for relating sequence and structural features to termination efficiency [83]. Similarly, by testing the effect of different hairpin structures on mRNA stabilities in multiple transcript contexts, it may be possible to identify rules to understand how the sequence and structure of a given hairpin affects half-life. Furthermore, as RNase III [34] or helicase activity [32] may mitigate the stabilizing effects of secondary structure, more study of RNA sequence and structure interactions with these enzymes should lead to better genetic design predictability.

10.2.3 Noncoding RNA-Mediated

ncRNA has been used for TSC in two related forms, namely, sRNA and asRNA. Both sRNA and asRNA act via an antisense mechanism and base-pair with a region – usually the 5′ UTR – of a target mRNA (Figure 10.2b). In one well-known example, asRNA was derived from the RNA-IN/RNA-OUT system from the insertion sequence IS10 in *E. coli* [84]. There, the RNA-IN antisense hairpin binds to the RNA-OUT portion of the target mRNA. Although engineered sRNA mechanisms have originated from distinct naturally occurring ncRNA systems, both asRNA [85] and sRNA target [74, 86] base pairing has been shown to be Hfq-mediated *in vitro*, so it is likely that asRNA and sRNA functions are mechanistically similar.

Substantial progress has been made in the past few years toward developing sRNA and noncoding *trans*-RNA as avenues for controlling gene expression. In 2011, Man *et al*. developed initial design principles for creating novel sRNAs with Hfq-binding sites and regions targeting enhanced green fluorescent protein (EGFP) and a native *E. coli* gene [10]. They tested 16 such sRNAs and reported relative expression knockdown levels ranging from 6% to 71%. Furthermore, they showed sRNA-dependent target mRNA half-life reduction and used a temperature-sensitive RNase E mutant to establish the RNase E dependence of target transcript level reduction. Surprisingly, the reduction in gene expression was unaffected by the presence or absence of RNase E, suggesting that sRNA binding alone was sufficient to reduce translation and that TSC does not play a dominant role in this system (see also [60]).

Sharma *et al*. randomized the antisense seed portion of the *E. coli* sRNA Spot42 [73] to screen for sRNAs that downregulate a natively targeted gene. After a single round of screening, sRNAs were identified that downregulate a

natively targeted gene with 45–145-fold repression – an improvement over the 27-fold repression of the native sRNA. After two rounds of screening for sRNAs against a gene with no native sRNA, sRNAs were identified that could repress the relative level of gene expression 23–85-fold.

Most recently, Ishikawa *et al.* [74], Park *et al.* [53], and Na *et al.* [75] have taken systematic approaches to uncover design principles for sRNAs that are effective at repressing gene expression. Ishikawa *et al.* studied the SgrS sRNA in *E. coli* using mutational analysis and Northern blotting to elucidate the Hfq-binding motif of that sRNA. This motif was incorporated into artificial sRNA against three mRNA targets that showed orthogonal Hfq-dependent knockdown via Northern blotting. The authors speculate that any mRNA can be effectively targeted by designing sRNAs with at least 14 nucleotides (nt) of sequence complementarity to the RBS and a *cis* Hfq-binding motif located within 10 nt. Na *et al.* screened native sRNA scaffolds and potential mRNA target binding sites around the SD region and found that a MicC sRNA scaffold with a binding site spanning the first 21 nt (not including the SD region) of the target mRNA was particularly effective. Using that insight, sRNAs were developed to target native genes in a microbial platform engineered to produce L-tyrosine and cadaverine. In both cases, the authors were able to employ sRNA-mediated genetic repression to divert metabolic flux and increase product formation in the engineered system. Collectively, these sRNA design studies suggest that an Hfq-binding, scaffold-based sRNA platform may provide a means of downregulating gene expression predictably, as binding energy of the antisense region is strongly correlated with repression capacity [75].

In addition to sRNA, at least two studies have utilized IS10-based asRNA against the 5′ UTR. The first study built a model from 529 possible combinations of 23 sense and antisense pairs (termed RNA-IN and RNA-OUT), which was then used to forward-engineer new regulators [76]. A second study built upon the RNA-IN and RNA-OUT system by adding a theophylline aptamer-based domain upstream of the RNA-IN asRNA, which functions similarly to a riboswitch in that gene expression is controlled through structures modulating ribosomal access to the RBS. Several designed mutants were screened to find an aptamer–RNA-IN pseudoknot interaction that impaired the RNA-IN asRNA's ability to bind its RNA-OUT partner when the aptamer domain was not bound to its ligand [87]. This provides another means of dynamic control and, given the similarities between asRNA and sRNA, indicates that sRNA could be engineered for dynamic control by appending ligand-binding aptamer domains.

10.2.4 Model-Driven Transcript Stability Control for Metabolic Pathway Engineering

Ribozyme-catalyzed phosphodiester bond cleavage can affect mRNA half-life in primarily two ways (Figure 10.3b). Depending on the sequence context, and whether the target site is within a 5′ or 3′ UTR, cleavage may remove or alter secondary structures that influence RNase E or ribosome docking, resulting in differences in transcript stability, and, potentially, levels of gene expression

[21, 78, 88]. Phosphodiester bond cleavage within the 5′ UTR also removes the 5′-P(PP) recognized by RppH or RNase E, which can lead to increased transcript persistence and gene expression [7, 21, 88] (see Section 10.1.2). We speculate that mRNA terminated with a 5′-OH is degraded via comparatively slow RNase E direct entry, consistent with increased half-lives measured for transcripts cleaved by hammerhead ribozymes [21].

Carothers *et al.* [7] formulated a model-driven process that uses UTR cleavage to engineer devices that regulate transcript stability and quantitatively program gene expression. Static ribozyme-regulated expression devices (rREDs) and dynamic, metabolite-controlled aptazyme-regulated expression devices (aREDs) were constructed that employ transcript stability, via 5′ UTR cleavage, as the underlying genetic control mechanism. With mechanistic understanding of RNA degradation pathways as a starting point [21, 88], a coarse-grained biochemical model of device function was created to simulate global device functions from local, measurable, and tunable component characteristics. The combinatorial space of design variable inputs was then mapped to the space of device outputs with a sampling-based approach, providing data for global sensitivity analysis (GSA) and identifying functional designs that meet targeted performance criteria. To physically implement functional devices, a novel method for designing transcripts with kinetic RNA folding simulations [89] was created that enables the assembly of individually characterized components parts.

To demonstrate that variations in tunable design parameters generate quantitatively predictable outputs, genetic devices were constructed to program amounts of a reporter protein and production levels of *p*-aminophenylalanine (p-AF), a chemical precursor of bioactive compounds and advanced polymers, from a 12-gene engineered biosynthetic pathway. In total, 28 *E. coli* expression devices were assembled from component parts that were generated and characterized separately *in vitro*, *in vivo*, and *in silico*. Excellent quantitative agreement between the design specifications and the device functions ($r = 0.94$) was observed, experimentally validating the underlying models and simulation tools and the overall approach. rREDs and aREDs have immediate utility as programmable biosensors and controllers for metabolic pathways and genetic circuits. And, notably, this work also provides a conceptual and experimental framework for investigating and engineering complex RNA functions through the application of fundamental biochemical understanding.

Using this framework, we envision a model-driven design process for creating RNA-based dynamic control systems for applications in metabolic engineering and biosensing (Figure 10.4). As a testbed for RNA-based control circuit design, we are engineering *E. coli* to produce *p*-aminostyrene (p-AS), a component of polymer composites with optical and mechanical properties favorable for advanced applications in photonics, photolithography, and biomedicine [90, 91]. Substituted styrenes have been difficult to chemically synthesize in high yields [92], and the cytotoxicity of key intermediates and products has prevented efficient microbial production [93]. The proposed *p*-AS pathway is an ideal testbed because it has 15 well-defined gene products and measurable intermediates yet presents a full complement of canonical control problems that must be addressed to obtain efficient production.

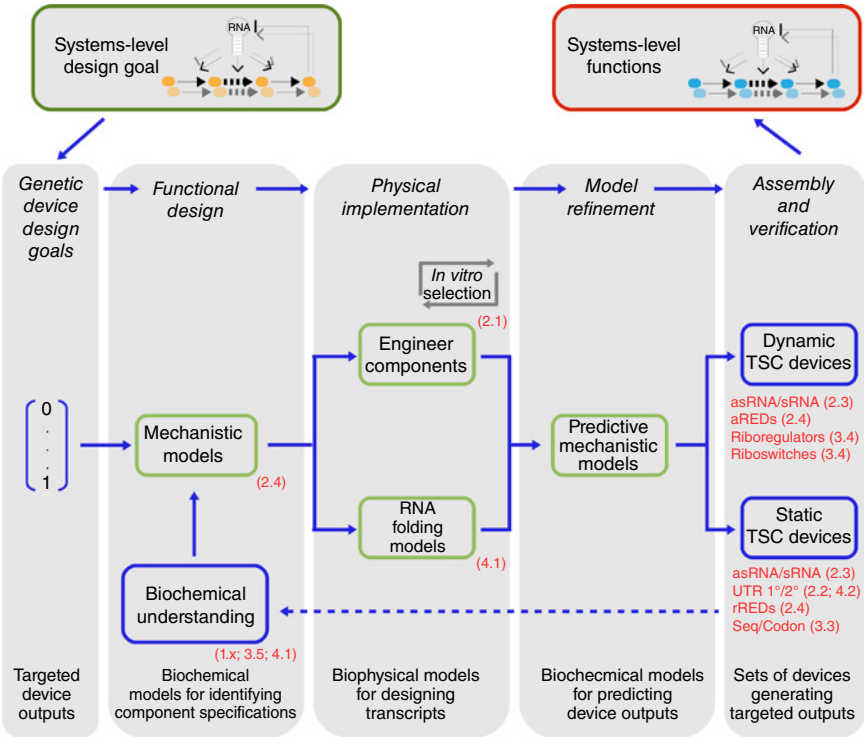


Figure 10.4 Model-driven design workflow for engineering gene expression with TSC. A basic systems-level model is used to identify goals for genetic device outputs. These goals inform the creation of a mechanistic model based on biochemical understanding, which is used to identify component specifications needed for device function. Components are then engineered and/or evolved with *in vitro* selection to meet the design specifications. Transcript design methods employing biophysical models of RNA folding are employed to enable the assembly of individual RNA components into functional devices. The mechanistic model is then refined to account for engineered component characteristics used to predict device outputs. Systems-level functions are obtained through the assembly of multiple static and dynamic RNA devices.

Drawing on the design principles gleaned from naturally occurring metabolic control circuits [94, 95], one approach to optimizing *p*-AS production would be to implement dynamic controllers that operate as a function of flux through *p*-AF, a cell-permeable intermediate. Circuits comprised of static rREDs and dynamic *p*-AF-responsive aREDs could be constructed to program flux through the pathway. In principle, there are many possible control topologies and corresponding RNA-based feedback architectures that could be implemented to enable high levels of *p*-AS production. An important aspect of this work will therefore be to identify and experimentally validate the feedback architectures that can be implemented across the tunable biochemical parameter ranges.

Finally, results showing the importance of robust folding to the design of functional rREDs and aREDs are consistent with the idea that kinetically driven co-transcriptional folding pathways significantly impact cellular RNAs [96].

Improving the ability to integrate biochemical models and refined RNA transcript folding design algorithms should therefore lead to better tools for engineering genetic control systems that employ RNA sequence and structure design to quantitatively program expression.

10.3 Managing Transcript Stability

10.3.1 Transcript Stability as a Confounding Factor

Perhaps the greatest obstacle on the road to predictable biological engineering is the joint confounding effect of cellular subsystems that interact with synthetic biological components in unanticipated ways. In this regard, it is important to realize that any – and all – synthetic RNA in the cell is affected by the degradation systems regulating transcript stability. Except in cases where transcript stability is the explicit genetic control mechanism, efforts aimed at engineering gene expression tend to neglect dimensions of transcript stability. However, as new tools are developed, it should become much easier to circumvent limitations imposed by variations in RNA stability and instead fine-tune transcript stability in concert with other engineering strategies to rapidly implement genetic controls to meet performance requirements.

10.3.2 Anticipating Transcript Stability Issues

Because so many factors can affect RNA stability, it is important to consider the ways that experimental results may be impacted by unexpected changes in transcript stability. Moreover, it may be prudent to routinely determine whether transcript stability is a parameter requiring attention, either through computational design variable sensitivity analysis or through wet-lab experimentation. The roles and binding behaviors of all ribonucleases and associated proteins have yet to be elucidated [97], but study of major players such as RNase E, PNPase, RppH, Hfq, and RNase III has unveiled structures and many key roles of these enzymes. Though the binding interactions of these enzymes with RNA and each other are not completely understood, it is possible to analyze sequences and attempt to avoid unwanted degradation, or increase degradation, by changing codons within the open reading frame or UTR sequences to eliminate, or insert, putative binding sites.

Computationally, the potential impact of transcript stability on a given synthetic biological device output can be assessed with GSA using coarse-grained mechanistic model simulations and Monte Carlo sampling [98]. With this method, the global space of potential designs is mapped by simulating genetic device outputs with Monte Carlo sampled values for the model parameters taken randomly from biochemically reasonable ranges. By computing quantitative GSA measures to relate the potential genetic device outputs to transcript stability parameter inputs (e.g., partial correlation coefficients) [98, 99], the impact of variations in RNA degradation rate, relative to other tunable design variables, can be readily discerned [7].

Considering transcript stability, and how it may impact system function, is especially important when introducing secondary structure into a transcript, as this may cause unwanted RNase binding or result in premature transcription termination. See Section 10.3.4 for an example detailing issues arising from adding *cis*-repressor riboregulator RNA into a 5′ UTR in *E. coli*. Another example comes from efforts to obtain detectable signals *in vivo* from RNA aptamer-based fluorescent biosensors (i.e., "Spinach" aptamer conjugates). To do this, Paige *et al*. had to employ an RNase E-deficient *E. coli* strain [100] to circumvent limitations likely resulting from an otherwise short aptamer half-life.

There are other potentially confounding effects that are more difficult to account for, but that should still be considered in the course of genetic device engineering. Large amounts of synthetic mRNA and/or regulatory RNA from complex circuit designs could lead to overloading the degradosome or associated enzymes such as Hfq, causing cell-wide RNA stability changes. A phenomenon of this sort is difficult to study, but the work of Hussein and Lim on competition for Hfq [49] suggests it is worth attempting to understand. They found that sRNA expressed without a target binding partner reduced sRNA effectiveness cell-wide by binding Hfq and limiting its accessibility. Expression of the target mRNA removed this problem, suggesting that balance in expressing synthetic sRNA can be critical.

10.3.3 Uniformity of 5′ and 3′ Ends

Variations in UTR sequence context may elicit differences in local secondary structure, which in turn may alter transcript stability and gene expression [78, 88]. One way to guard against such context-dependent transcript stability problems is to attenuate UTR variability effects by removing 5′ and/or 3′ RNA that may form undesired secondary structure (Figure 10.3b). Several studies have utilized removal of 5′ UTR secondary structures as a mechanism for minimizing context-dependent differences in gene expression. One involved a screen of "insulator" sequences and structures placed within the 5′ UTR. A ribozyme–hairpin combination, termed RiboJ, produced nearly identical transfer functions for two different genes under the control of three different promoters (several other ribozymes had similar effects) [78]. A second study used the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) processing system from *Pseudomonas aeruginosa* strain UCBPP-PA14 to remove both 5′ and 3′ UTR sequences. At both ends, a 28-nucleotide repetitive sequence, recognized by the Csy4 endonuclease, was added, which resulted in efficient transcript cleavage and UTR sequence removal. Using the CRISPR system, they were able to show similar levels of protein production in the context of different promoter and RBS combinations in mono- and bicistronic systems, with green fluorescent protein (GFP) and red fluorescent protein (RFP) outputs [88].

Mutalik *et al*. recently published a scheme for minimizing 5′ UTR-induced variations in gene expression that involves introducing a standby RBS in a bicistronic design (BCD) [101]. The standby RBS is designed to cause ribosome binding upstream of the real RBS (i.e., the RBS from which translation of the desired

open reading frame predominantly initiates), disrupting secondary structure that may cause variability in ribosome binding to the real RBS. The use of a BCD resulted in a large improvement of transfer function correlations between multiple gene contexts.

With the three individual systems in this section as starting points, it may be possible to combine them into a general-use template for attenuating 5′ UTR- and 3′ UTR-induced variations in gene expression. Further reductions in coding sequence-induced variability could come from work that identifies sequences and structures that, through targeted codon changes, minimize RNase E direct entry and RNase III binding.

10.3.4 RBS Sequestration by Riboregulators and Riboswitches

As mentioned in the Introduction, RBS sequestration can hasten mRNA degradation by leaving the transcript open to binding from RNases. Thus, riboregulators, riboswitches, and similar structures that sequester the RBS can be used to dynamically control mRNA stability, at least in part, via induction (or repression) by a *trans*-activating RNA or small molecule (Figure 10.3c). Riboregulators are functional RNA structures with two components: a *cis*-repressor and a *trans*-activator. The *cis*-repressor is generally 5′ UTR RNA that folds into a conformation that base-pairs with the RBS, making it unavailable to ribosomes. This repression can be relieved by the *trans*-activator RNA, an RNA that interacts with the *cis*-repressor RNA so that the RBS is revealed for translation [54]. The mechanistic design can also be reversed, whereby *trans*-RNA binding changes the *cis*-RNA conformation to sequester the RBS [54]. Like riboregulators, riboswitches can function to bind the RBS in *cis* and prevent or enhance translation by hindering, or allowing, ribosome docking. In the place of *trans*-RNA, riboswitches control RBS sequestration with conformational changes mediated by small molecule ligand binding. Functional synthetic riboswitches have been developed to bind a variety of ligands [102–105].

As both riboregulators and riboswitches involve adding secondary structure into the 5′ UTR, their presence is likely to cause altered transcript stability due to changes in RNase binding site accessibility. It may prove useful to think of these systems in terms of their dual effects on translation rate and transcript degradation rate, which will require gathering data related to half-life, and not just final gene expression and protein output. Both sets of data would be necessary to decouple the contributions of translation rate changes from transcript stability changes.

Efforts by the Collins lab to engineer synthetic riboregulators show how transcript stability changes can impact device outputs. The addition of *cis*-repressor RNA in the 5′ UTR of a particular GFP expression cassette significantly reduced protein expression, and repression was largely alleviated by the *trans*-activator RNA [9]. The synthetic riboregulator system was subsequently expanded to develop a microbial kill switch [106] and a genetic switchboard to regulate four carbon-utilization genes [107]. Though this system has been utilized successfully, it is worth noting that inserting *cis*-repressor RNA into their constructs led to a 40% reduction in mRNA levels versus no *cis*-repressor RNA [9], which the

authors attribute to either premature transcript termination due to the *cis*-repressor secondary structure or the activities of RNases, such as RNase III, that cleave double-stranded RNA.

Elsewhere, the space of *trans*-activating RNA targeting a fixed *cis*-repressor RNA, regulating GFP expression in *E. coli*, has been computationally explored [108]. In the *cis*-repressed state, where the level of genetic output was equivalent to 1–4% of the unrepressed state, activation by one of six designed *trans*-activating RNAs increased GFP production 3–11-fold relative to the baseline. Quantitative reverse transcription polymerase chain reaction (RT-PCR) showed that the *trans*-activating RNA–mRNA ratio did not change in an RNase III knockout strain compared with the wild type. However, the relative genetic output induced by *trans*-activating RNA more than doubled in the RNase III knockout, which, as mentioned, is consistent with the idea that variations in transcript stability can alter the performance characteristics of these kinds of control devices and systems.

A riboregulator-like RNA, called an allosteric ribozyme, previously only characterized *in vitro* [109–111], has recently been used to control translation initiation *in vivo* [112]. A ribozyme, with the RBS sequestered in its secondary structure, was designed to autocatalytically cleave itself to expose the RBS and allow translation initiation. *Trans*-activating RNAs were designed to bind a complementary sequence within the ribozyme, inhibiting ribozyme cleavage and exposure of the RBS, leading to 10-fold reductions in relative EGFP expression. The question of how variations in transcript stability might be in play here has not been directly investigated.

Overall, the work described here highlights promising approaches for engineering dynamic RNA-based control systems. It is also clear that, to improve engineering tractability, there is a need to investigate how introducing secondary structure – whether a *cis*-repressor RNA, riboswitch, or ribozyme – into the 5′ UTR may lead to confounding effects on device outputs stemming from unaccounted-for RNase binding or premature transcription termination.

10.3.5 Experimentally Probing Transcript Stability

The determinants of synthetic transcript stability can be analyzed by experimentally measuring mRNA half-life, through expression studies, by the use of endonuclease gene knockout strains, and with computational RNA folding simulations. Quantitative gel electrophoresis [21], quantitative PCR, or RNA-seq [1] after the addition of a transcription-inhibiting antibiotic (e.g., rifampicin) can be done at intervals to determine average transcript half-life by quantifying transcript abundance as a function of time. A strategy using sRNA to quantify mRNA abundance changes has also been proposed [113]. Comparing measurements from cells with and without an RNase (via knockout) can lend insight into the RNase dependence of a phenomenon, though care should be taken to understand the global impact of an RNase deletion and how that may complicate data interpretation. Folding simulation tools for calculating minimum free energy (MFE) secondary structures [114–117] or kinetically driven co-transcriptional [118–120] folding trajectories can lend insight into whether secondary structure could be

problematic (see Section 10.4.1 for more information). As more researchers use these tools to better understand synthetic RNA function, the aggregated data will lead to further insights and design strategies that can take advantage of stability or mitigate unwanted effects.

10.4 Potential Mechanisms for Transcript Control

10.4.1 Leveraging New Tools

The advent of recent technologies, such as high-throughput RNA secondary structure elucidation, high-throughput RNA sequencing, and co-transcriptional RNA folding simulations, provides new ways to investigate transcript control for predictable gene expression engineering. Most artificial RNA-based control strategies have yet to take full advantage of these technologies to more predictably engineer synthetic systems.

High-throughput RNA structural sequencing [121, 122] presents a new way to examine the structures of large numbers of RNA molecules. These techniques use RNA cleavage events dependent on the absence or presence of secondary structure, followed by high-throughput sequencing, to develop a map of base pairing probabilities. This map can then be used to constrain models from structure prediction software [123–125]. If paired with half-life quantification, this methodology could provide better understanding of the connection between UTR structure and transcript stability, enabling more predictable introduction of secondary structure into transcripts.

MFE simulations using tools like Mfold [114, 115] or RNAfold [116, 117] have been a mainstay in RNA secondary structure prediction. These tools calculate the lowest energy state of an RNA, which is interpreted to be the steady-state conformation of the RNA. When attempting to predict RNA secondary structures inside cells, MFE calculations may be misleading, for at least two reasons. First, mRNA folding is co-transcriptional, and thus the full transcript sequence is not available for folding at all times. Second, the relatively short half-life of most mRNAs in the cell [2] can preclude their reaching the MFE conformation before degradation. To address these issues, there are software packages that take co-transcriptional effects into account [118–120] and can thus be useful for predicting UTR secondary structures more accurately in a cellular context. In fact, the creation of a transcript design method built around kinetic co-transcriptional RNA folding simulations was crucial for the rRED and aRED engineering described in Section 10.2.4 [7]. In that work, custom software written to implement kinefold [119] on a computational cluster enabled the design of spacer sequences to allow assembly of individually generated and characterized RNA parts into genetic devices with quantitatively predictable functions. There was significant divergence between the transcript folds predicted with MFE structure calculations and those obtained with kinetic simulations, underscoring the importance of RNA sequence and structure design that explicitly considers co-transcriptional folding. To extend those results, we are currently developing a computational platform for designing RNA parts, devices, and transcripts with

kinetic folding algorithms that should be broadly useful for analyzing and engineering TSC mechanisms [89].

10.4.2 Unused Mechanisms Found in Nature

Despite more than a decade of progress, there are naturally occurring mechanisms for controlling transcript stability that have yet to be exploited for engineering gene expression. In the following, we enumerate several promising mechanisms that could become part of a toolkit to predictably control transcript stability.

The *lysC* riboswitch, recently described by Caron *et al.* (Figure 10.5b) [28], contains two RNase E cleavage sites that are exposed, while the RBS is simultaneously sequestered, in the presence of the ligand lysine. In a synthetic context, rational introduction of these sites into existing riboswitch designs could enhance their function by decreasing transcript half-life in the presence of ligand.

An sRNA titration system has been observed by two different groups in *Salmonella* (Figure 10.5a) [126, 127]. This system functions by relieving sRNA-based repression by expressing a decoy mRNA to shunt away sRNA targeting an mRNA, leading to rapid degradation of the sRNA. Such a system could allow for quick removal of engineered sRNA-based repression that rapidly activates gene expression.

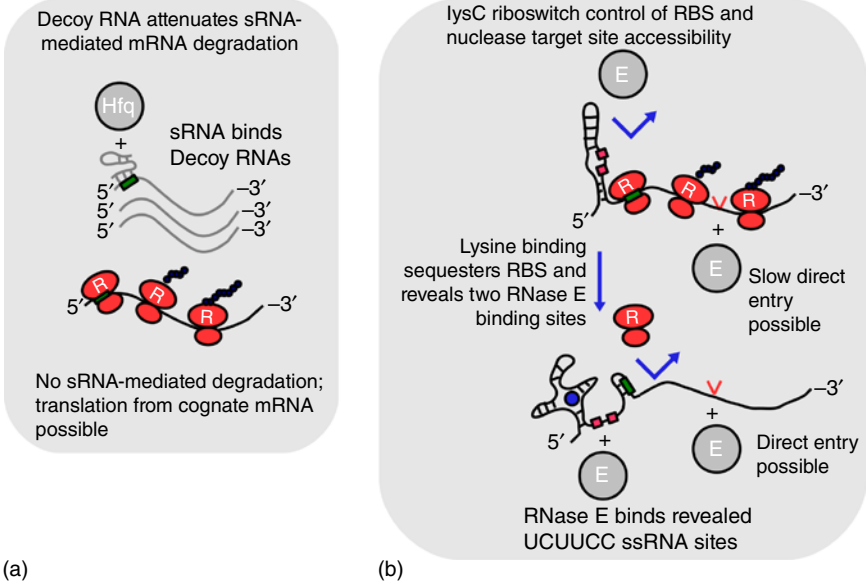


Figure 10.5 Examples of unutilized transcript stability control mechanisms. (a) In naturally occurring biological systems, decoy RNAs attenuate sRNA-mediated mRNA degradation by shunting sRNAs away from cognate transcripts, increasing message stability and the level of genetic expression. (b) The *lysC* transcript of *E. coli* contains a riboswitch that controls access to the RBS and RNase E target sites, showing the functional integration of multiple TSC mechanisms.

RNA–protein interactions occur throughout the lifetime of a transcript and play critical roles in the degradation process. As results with the bicistronic 5′ UTR design [101] highlighted in Section 10.3.3 suggest, some RNA–protein interactions can effectively inhibit other RNA–protein interactions. Although this cross-inhibition of RNA–protein interactions sometimes confounds system behavior, this principle could be exploited as a TSC mechanism. For example, sequence motifs that recruit protective RNA-binding proteins to the 5′ and 3′ UTRs could insulate transcripts from ribonuclease binding. Pentatricopeptide repeat (PPR) proteins, a family of single-stranded RNA-binding proteins found in plants [128], would be a good candidate for this application. PPR proteins are similar to the DNA-binding transcription activator-like effector (TALE) proteins [129] in that each protein has an RNA-binding domain, the target specificity of which is governed by a series of two-amino-acid repeats, where each repeat corresponds to a target nucleotide. The sequence motif code with which a class of PPR proteins binds RNA has recently been uncovered [130]. PPR proteins have been implicated in controlling transcript stability in maize chloroplasts by binding to the 5′ and 3′ ends of mRNA [131], and a PPR protein was shown to limit $5' \rightarrow 3'$ and $3' \rightarrow 5'$ degradation *in vitro* when its binding site was introduced into an mRNA [132].

Similarly to the PPR proteins, RNA has been found to bind the 3′ UTR of a transcript and enhance its stability by offering protection against exoribonuclease activity. GadY is an asRNA with complementarity to the 3′ UTR of the *gadX* gene in *E. coli* [56] and is probably one of many such asRNA.

Though the biochemical details involving polyadenylation are not yet fully understood in bacteria, its ubiquity [133] and use in contexts such as in the *glmS* gene [63] for increasing degradation highlight potential utility for engineering TSC. As *E. coli* has only $3' \rightarrow 5'$ exoribonucleases, degradation from the 3′ end is an essential part of rendering a transcript nonfunctional. Adding poly(A) tails of varying lengths – perhaps in an inducible manner similar to a riboswitch or riboregulator – to transcripts could function to reliably control and enhance 3′ end degradation by PNPase.

10.5 Conclusions and Discussion

Knowledge of RNA degradation in bacteria has progressed substantially since the advent of synthetic biology. Key components and processes that account for bulk mRNA turnover, translation effects, sRNA action, and polyadenylation have become well understood. With this know-how and the continuing efforts of the RNA-based engineering community, TSC is positioned to become an even more powerful method for programming functions in synthetic biological systems. A forward-engineering approach that harnesses understanding of biochemical mechanisms to build predictive models for generating desired outputs [7] is now possible, with a number of mechanisms to up- and downregulate transcript half-life. Building on existing RNA device engineering efforts, inspiration from natural mechanisms can point to new ways of regulating stability; and as RNA device engineering matures, more complex and wholly synthetic devices

that, for instance, fuse multiple control mechanisms will become easier to design and construct. Moreover, increasing knowledge of the underlying biochemical and biophysical principles governing RNA degradation will make it easier to anticipate how transcript stability may function contrary to genetic device output goals. We expect that TSC engineering will create ways to align mRNA half-life with design goals and thus will be a vital component of increasing system predictability and avoiding confounding effects. The current state of research portends the use of TSC to help design synthetic biological systems that can dynamically and rapidly respond to their environment with low-burden, orthogonal RNA components designed entirely *in silico*.

Acknowledgments

We thank W.M. Voje, Jr. and R.C. Correa for helpful discussions and comments. Work in the authors' laboratory is supported by the Molecular Engineering & Sciences Institute, the University of Washington, and the National Science Foundation Synthetic Biology Engineering Research Center (NSF SynBERC). J.M.C. is a fellow of the Alfred P. Sloan Foundation.

Definitions

RNA synthetic biology Large-scale genetic engineering that makes use of RNA-based components (e.g., aptamers, aptazymes, riboswitches) to construct control devices and systems for programming cellular function

Transcript stability control Regulation of genetic output by engineering mRNA degradation rate

mRNA degradation The process by which a transcript is hydrolyzed to component monomers by the concerted efforts of the degradosome and associated enzymes

Aptamer Functional RNA structure, typically generated through *in vitro* selection, that binds target ligands

Ribozyme Functional RNA structure with catalytic activity (e.g., hammerhead ribozymes catalyze phosphodiester bond cleavage reactions)

Aptazyme Composite functional RNA structure consisting of an aptamer and a phosphodiester bond-cleaving ribozyme such that the catalytic activity is modulated by aptamer ligand binding

Riboswitch RNA structure that controls gene expression by employing a ligand-binding aptamer domain that regulates access of the ribosome to the ribosome binding site in *cis*

Riboregulator A riboswitch-like RNA unit that regulates access to the ribosome binding site in response to binding by a *trans*-RNA

sRNA Small RNAs, usually tens to hundreds of nucleotides long, that bind regions of a target mRNA (typically near the 5′ UTR or start codon) to hasten mRNA degradation and prevent translation by occluding ribosome docking

Computational design Here, a methodology that utilizes biochemical and biophysical models to drive the construction of complex devices and systems with predictable functions

References

1 Kristoffersen, S.M., Haase, C., Weil, M.R., Passalacqua, K.D. *et al.* (2012) Global mRNA decay analysis at single nucleotide resolution reveals segmental and positional degradation patterns in a Gram-positive bacterium. *Genome Biol.*, **13**, R30.

2 Selinger, D.W., Saxena, R.M., Cheung, K.J., Church, G.M. *et al.* (2003) Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. *Genome Res.*, **13**, 216–223.

3 Bandyra, K.J., Bouvier, M., Carpousis, A.J., and Luisi, B.F. (2013) The social fabric of the RNA degradosome. *Biochim. Biophys. Acta*, **1829**, 514–522.

4 Anderson, K.L. and Dunman, P.M. (2009) Messenger RNA turnover processes in *Escherichia coli*, *Bacillus subtilis*, and emerging studies in *Staphylococcus aureus*. *Int. J. Microbiol.*, **2009**, 525491.

5 Steege, D.A. (2000) Emerging features of mRNA decay in bacteria. *RNA (New York, N.Y.)*, **6**, 1079–1090.

6 Desnoyers, G., Bouchard, M.-P., and Massé, E. (2013) New insights into small RNA-dependent translational regulation in prokaryotes. *Trends Genet.*, **29**, 92–98.

7 Carothers, J.M., Goler, J.A., Juminaga, D., and Keasling, J.D. (2011) Model-driven engineering of RNA devices to quantitatively program gene expression. *Science (New York, N.Y.)*, **334**, 1716–1719.

8 Carrier, T.A. and Keasling, J.D. (1997) Controlling messenger RNA stability in bacteria: strategies for engineering gene expression. *Biotechnol. Progr.*, **13**, 699–708.

9 Isaacs, F.J., Dwyer, D.J., Ding, C., Pervouchine, D.D. *et al.* (2004) Engineered riboregulators enable post-transcriptional control of gene expression. *Nat. Biotechnol.*, **22**, 841–847.

10 Man, S., Cheng, R., Miao, C., Gong, Q. *et al.* (2011) Artificial trans-encoded small non-coding RNAs specifically silence the selected gene expression in bacteria. *Nucleic Acids Res.*, **39**, e50.

11 Holtz, W.J. and Keasling, J.D. (2010) Engineering static and dynamic control of synthetic pathways. *Cell*, **140**, 19–23.

12 Górna, M.W., Carpousis, A.J., and Luisi, B.F. (2012) From conformational chaos to robust regulation: the structure and function of the multi-enzyme RNA degradosome. *Q. Rev. Biophys.*, **45**, 105–145.

13 Carpousis, A.J. (2007) The RNA degradosome of *Escherichia coli* : an mRNA-degrading machine assembled on RNase E. *Annu. Rev. Microbiol.*, **61**, 71–87.

14 Kaberdin, V.R. and Lin-chao, S. (2009) Unraveling new roles for minor components of the *E. coli* RNA degradosome. *RNA Biol.*, **6**, 402–405.

15 Taghbalout, A. and Rothfield, L. (2007) RNaseE and the other constituents of the RNA degradosome are components of the bacterial cytoskeleton. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 1667–1672.

16 Khemici, V., Poljak, L., Luisi, B.F., and Carpousis, A.J. (2008) The RNase E of *Escherichia coli* is a membrane-binding protein. *Mol. Microbiol.*, **70**, 799–813.

17 Mackie, G.A. (2013) RNase E: at the interface of bacterial RNA processing and decay. *Nat. Rev. Microbiol.*, **11**, 45–57.

18 Bouvet, P. and Belasco, J.G. (1992) Control of RNase E-mediated RNA degradation by 5′-terminal base pairing in *E. coli*. *Nature*, **360**, 488–491.

19 Mackie, G.A. (1998) Ribonuclease E is a 5′-end-dependent endonuclease. *Nature*, **395**, 720–723.

20 Callaghan, A.J., Marcaida, M.J., Stead, J.A., McDowall, K.J. *et al.* (2005) Structure of *Escherichia coli* RNase E catalytic domain and implications for RNA turnover. *Nature*, **437**, 1187–1191.

21 Celesnik, H., Deana, A., and Belasco, J.G. (2007) Initiation of RNA decay in *Escherichia coli* by 5′ pyrophosphate removal. *Mol. cell*, **27**, 79–90.

22 Deana, A., Celesnik, H., and Belasco, J.G. (2008) The bacterial enzyme RppH triggers messenger RNA degradation by 5′ pyrophosphate removal. *Nature*, **451**, 355–358.

23 Bouvier, M. and Carpousis, A.J. (2011) A tale of two mRNA degradation pathways mediated by RNase E. *Mol. Microbiol.*, **82**, 1305–1310.

24 Joyce, S.A. and Dreyfus, M. (1998) In the absence of translation, RNase E can bypass 5′ mRNA stabilizers in *Escherichia coli*. *J. Mol. Biol.*, **282**, 241–254.

25 Baker, K.E. and Mackie, G.A. (2003) Ectopic RNase E sites promote bypass of 5′-end-dependent mRNA decay in *Escherichia coli*. *Mol. Microbiol.*, **47**, 75–88.

26 Kime, L., Jourdan, S.S., Stead, J.A., Hidalgo-Sastre, A. *et al.* (2010) Rapid cleavage of RNA by RNase E in the absence of 5′ monophosphate stimulation. *Mol. Microbiol.*, **76**, 590–604.

27 Kaberdin, V.R. (2003) Probing the substrate specificity of *Escherichia coli* RNase E using a novel oligonucleotide-based assay. *Nucleic Acids Res.*, **31**, 4710–4716.

28 Caron, M.-P., Bastet, L., Lussier, A., Simoneau-Roy, M. *et al.* (2012) Dual-acting riboswitch control of translation initiation and mRNA decay. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, E3444–E3453.

29 Vanzo, N.F., Li, Y.S., Py, B., Blum, E. *et al.* (1998) Ribonuclease E organizes the protein interactions in the *Escherichia coli* RNA degradosome. *Genes Dev.*, **12**, 2770–2781.

30 Braun, F., Hajnsdorf, E., and Regnier, P. (1996) Polynucleotide phosphorylase is required for the rapid degradation of the RNase E-processed rpsO mRNA of *Escherichia coli* devoid of its 3′ hairpin. *Mol. Microbiol.*, **19**, 997–1005.

31 Py, B., Higgins, C.F., Krisch, H.M., and Carpousis, A.J. (1996) A DEAD-box RNA helicase in the *Escherichia coli* RNA degradosome. *Nature*, **381**, 169–172.

32 Khemici, V., Poljak, L., Toesca, I., and Carpousis, A.J. (2005) Evidence in vivo that the DEAD-box RNA helicase RhlB facilitates the degradation of ribosome-free mRNA by RNase E. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 6913–6918.

33 Morita, T., Kawamoto, H., Mizota, T., Inada, T. *et al.* (2004) Enolase in the RNA degradosome plays a crucial role in the rapid decay of glucose transporter

mRNA in the response to phosphosugar stress in *Escherichia coli*. *Mol. Microbiol.*, **54**, 1063–1075.

34 Pertzev, A.V. and Nicholson, A.W. (2006) Characterization of RNA sequence determinants and antideterminants of processing reactivity for a minimal substrate of *Escherichia coli* ribonuclease III. *Nucleic Acids Res.*, **34**, 3708–3721.

35 MacRae, I.J. and Doudna, J.A. (2007) Ribonuclease revisited: structural insights into ribonuclease III family enzymes. *Curr. Opin. Struct. Biol.*, **17**, 138–145.

36 Ow, M.C., Perwez, T., and Kushner, S.R. (2003) RNase G of *Escherichia coli* exhibits only limited functional overlap with its essential homologue, RNase E. *Mol. Microbiol.*, **49**, 607–622.

37 Deana, A. and Belasco, J.G. (2005) Lost in translation: the influence of ribosomes on bacterial mRNA decay. *Genes Dev.*, **19**, 2526–2533.

38 Vytvytska, O., Moll, I., Kaberdin, V.R., von Gabain, A. *et al.* (2000) Hfq (HF1) stimulates ompA mRNA decay by interfering with ribosome binding. *Genes Dev.*, **14**, 1109–1118.

39 Komarova, A.V., Tchufistova, L.S., Dreyfus, M., and Boni, I.V. (2005) AU-rich sequences within 5′ untranslated leaders enhance translation and stabilize mRNA in *Escherichia coli*. *J. Bacteriol.*, **187**, 1344–1349.

40 Makarova, O.V., Makarov, E.M., Sousa, R., and Dreyfus, M. (1995) Transcribing of *Escherichia coli* genes with mutant T7 RNA polymerases: stability of lacZ mRNA inversely correlates with polymerase speed. *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 12250–12254.

41 Braun, F., Le Derout, J., and Régnier, P. (1998) Ribosomes inhibit an RNase E cleavage which induces the decay of the rpsO mRNA of *Escherichia coli*. *EMBO J.*, **17**, 4790–4797.

42 Smith, A.M., Fuchs, R.T., Grundy, F.J., and Henkin, T.M. (2010) Riboswitch RNAs: regulation of gene expression by direct monitoring of a physiological signal. *RNA Biol.*, **7**, 104–110.

43 Sharma, C.M. and Vogel, J. (2009) Experimental approaches for the discovery and characterization of regulatory small RNA. *Curr. Opin. Microbiol.*, **12**, 536–546.

44 Storz, G., Vogel, J., and Wassarman, K.M. (2011) Regulation by small RNAs in bacteria: expanding frontiers. *Mol. cell*, **43**, 880–891.

45 Aiba, H. (2007) Mechanism of RNA silencing by Hfq-binding small RNAs. *Curr. Opin. Microbiol.*, **10**, 134–139.

46 Vogel, J. and Luisi, B.F. (2011) Hfq and its constellation of RNA. *Nat. Rev. Microbiol.*, **9**, 578–589.

47 Raghavan, R., Groisman, E.a., and Ochman, H. (2011) Genome-wide detection of novel regulatory RNAs in *E. coli*. *Genome Res.*, **21**, 1487–1497.

48 Bandyra, K.J., Said, N., Pfeiffer, V., Górna, M.W. *et al.* (2012) The seed region of a small RNA drives the controlled destruction of the target mRNA by the endoribonuclease RNase E. *Mol. Cell*, **47**, 943–953.

49 Hussein, R. and Lim, H.N. (2011) Disruption of small RNA signaling caused by competition for Hfq. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 1110–1115.

50 Andrade, J.M., Pobre, V., Matos, A.M., and Arraiano, C.M. (2012) The crucial role of PNPase in the degradation of small RNAs that are not associated with Hfq. *RNA (New York, N.Y.)*, **18**, 844–855.

51 Sittka, A., Lucchini, S., Papenfort, K., Sharma, C.M. *et al.* (2008) Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLos Genet.*, **4**, e1000163.

52 Zhang, A., Wassarman, K.M., Rosenow, C., Tjaden, B.C. *et al.* (2003) Global analysis of small RNA and mRNA targets of Hfq. *Mol. Microbiol.*, **50**, 1111–1124.

53 Park, H., Bak, G., Kim, S.C., and Lee, Y. (2013) Exploring sRNA-mediated gene silencing mechanisms using artificial small RNAs derived from a natural RNA scaffold in *Escherichia coli*. *Nucleic Acids Res.*, **41**, 3787–3804.

54 Lease, R.A., Cusick, M.E., and Belfort, M. (1998) Riboregulation in *Escherichia coli*: DsrA RNA acts by RNA:RNA interactions at multiple loci. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 12456–12461.

55 Majdalani, N., Cunning, C., Sledjeski, D., Elliott, T. *et al.* (1998) DsrA RNA regulates translation of RpoS message by an anti-antisense mechanism, independent of its action as an antisilencer of transcription. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 12462–12467.

56 Opdyke, J.A., Kang, J., and Storz, G. (2004) GadY, a small-RNA regulator of acid response genes in *Escherichia coli*. *J. Bacteriol.*, **186**, 6698–6705.

57 Antal, M., Bordeau, V., Douchin, V., and Felden, B. (2005) A small bacterial RNA regulates a putative ABC transporter. *J. Biol. Chem.*, **280**, 7901–7908.

58 Chen, S., Zhang, A., Blyn, L.B., and Storz, G. (2004) MicC, a second small-RNA regulator of Omp protein expression in *Escherichia coli*. *J. Bacteriol.*, **186**, 6689–6697.

59 Vanderpool, C.K. and Gottesman, S. (2004) Involvement of a novel transcriptional activator and small RNA in post-transcriptional regulation of the glucose phosphoenolpyruvate phosphotransferase system. *Mol. Microbiol.*, **54**, 1076–1089.

60 Morita, T., Mochizuki, Y., and Aiba, H. (2006) Translational repression is sufficient for gene silencing by bacterial small noncoding RNAs in the absence of mRNA destruction. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 4858–4863.

61 Li, Z., Reimers, S., Pandit, S., and Deutscher, M.P. (2002) RNA quality control: degradation of defective transfer RNA. *EMBO J.*, **21**, 1132–1138.

62 Andrade, J.M., Hajnsdorf, E., Régnier, P., and Arraiano, C.M. (2009) The poly(A)-dependent degradation pathway of rpsO mRNA is primarily mediated by RNase R. *RNA (New York, N.Y.)*, **15**, 316–326.

63 Joanny, G., Le Derout, J., Bréchemier-Baey, D., Labas, V. *et al.* (2007) Polyadenylation of a functional mRNA controls gene expression in *Escherichia coli*. *Nucleic Acids Res.*, **35**, 2494–2502.

64 Hajnsdorf, E., Braun, F., Haugel-Nielsen, J., and Régnier, P. (1995) Polyadenylylation destabilizes the rpsO mRNA of *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 3973–3977.

65 Cao, G.J. and Sarkar, N. (1992) Identification of the gene for an *Escherichia coli* poly(A) polymerase. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 10380–10384.

66 O'Hara, E.B., Chekanova, J.a., Ingle, C.A., Kushner, Z.R. *et al.* (1995) Polyadenylylation helps regulate mRNA decay in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 1807–1811.

67 Carrier, T.A. and Keasling, J.D. (1997) Engineering mRNA stability in *E. coli* by the addition of synthetic hairpins using a 5′ cassette system. *Biotechnol. Bioeng.*, **55**, 577–580.

68 Chen, L.H., Emory, S.a., Bricker, A.L., Bouvet, P. *et al.* (1991) Structure and function of a bacterial mRNA stabilizer: analysis of the 5′ untranslated region of ompA mRNA. *J. Bacteriol.*, **173**, 4578–4586.

69 Lopez, P.J. and Dreyfus, M. (1996) The lacZ mRNA can be stabilised by the T7 late mRNA leader in *E coli*. *Biochimie*, **78**, 408–415.

70 Blum, E. (1999) Polyadenylation promotes degradation of 3′-structured RNA by the *Escherichia coli* mRNA degradosome *in vitro*. *J. Biol. Chem.*, **274**, 4009–4016.

71 Wong, H.C. and Chang, S. (1986) Identification of a positive retroregulator that stabilizes mRNAs in bacteria. *Proc. Natl. Acad. Sci. U.S.A.*, **83**, 3233–3237.

72 Smolke, C.D., Carrier, T.A., and Keasling, J.D. (2000) Coordinated, differential expression of two genes through directed mRNA cleavage and stabilization by secondary structures. *Appl. Environ. Microbiol.*, **66**, 5399–5405.

73 Sharma, V., Yamamura, A., and Yokobayashi, Y. (2012) Engineering artificial small RNAs for conditional gene silencing in *Escherichia coli*. *ACS Synth. Biol.*, **1**, 6–13.

74 Ishikawa, H., Otaka, H., Maki, K., Morita, T. *et al.* (2012) The functional Hfq-binding module of bacterial sRNAs consists of a double or single hairpin preceded by a U-rich sequence and followed by a 3′ poly(U) tail. *RNA (New York, N.Y.)*, **18**, 1062–1074.

75 Na, D., Yoo, S.M., Chung, H., Park, H. *et al.* (2013) Metabolic engineering of *Escherichia coli* using synthetic small regulatory RNAs. *Nat. Biotechnol.*, **31**, 170–174.

76 Mutalik, V.K., Qi, L., Guimaraes, J.C., Lucks, J.B. *et al.* (2012) Rationally designed families of orthogonal RNA regulators of translation. *Nat. Chem. Biol.*, **8**, 447–454.

77 Lucks, J.B., Qi, L., Mutalik, V.K., Wang, D. *et al.* (2011) Versatile RNA-sensing transcriptional regulators for engineering genetic networks. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 8617–8622.

78 Lou, C., Stanton, B., Chen, Y.-J., Munsky, B. *et al.* (2012) Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nat. Biotechnol.*, **30**, 1137–1142.

79 Carothers, J.M., Goler, J.A., Kapoor, Y., Lara, L. *et al.* (2010) Selecting RNA aptamers for synthetic biology: investigating magnesium dependence and predicting binding affinity. *Nucleic Acids Res.*, **38**, 2736–2747.

80 Lynch, S.A., Desai, S.K., Sajja, H.K., and Gallivan, J.P. (2007) A high-throughput screen for synthetic riboswitches reveals mechanistic insights into their function. *Chem. Biol.*, **14**, 173–184.

81 Carrier, T.A. and Keasling, J.D. (1999) Library of synthetic 5′ secondary structures to manipulate mRNA stability in *Escherichia coli*. *Biotechnol. Progr.*, **15**, 58–64.

82 Pfleger, B.F., Pitera, D.J., Smolke, C.D., and Keasling, J.D. (2006) Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. *Nat. Biotechnol.*, **24**, 1027–1032.

83 Cambray, G., Guimaraes, J.C., Mutalik, V.K., Lam, C. *et al.* (2013) Measurement and modeling of intrinsic transcription terminators. *Nucleic Acids Res.*, 1–10.

84 Kittle, J.D., Simons, R.W., Lee, J., and Kleckner, N. (1989) Insertion sequence IS10 anti-sense pairing initiates by an interaction between the 5′ end of the target RNA and a loop in the anti-sense RNA. *J. Mol. Biol.*, **210**, 561–572.

85 Ross, J.A., Ellis, M.J., Hossain, S., and Haniford, D.B. (2013) Hfq restructures RNA-IN and RNA-OUT and facilitates antisense pairing in the Tn10/IS10 system. *RNA (New York, N.Y.)*, **19**, 670–684.

86 Møller, T., Franch, T., Højrup, P., Keene, D.R. *et al.* (2002) Hfq: a bacterial Sm-like protein that mediates RNA–RNA interaction. *Mol. Cell*, **9**, 23–30.

87 Qi, L., Lucks, J.B., Liu, C.C., Mutalik, V.K. *et al.* (2012) Engineering naturally occurring trans-acting non-coding RNAs to sense molecular signals. *Nucleic Acids Res.*, **40**, 5775–5786.

88 Qi, L., Haurwitz, R.E., Shao, W., Doudna, J.A. *et al.* (2012) RNA processing enables predictable programming of gene expression. *Nat. Biotechnol.*, **30**, 1002–1006.

89 Thimmaiah, T., Voje, W.E. Jr., and Carothers, J.M. (2015) Computational design of RNA parts, devices, and transcripts with kinetic folding algorithms implemented on multiprocessor clusters. *Methods Mol. Biol.*, **1244**, 45–61.

90 Goikhman, M.Y., Yevlampieva, N.P., Kamanina, N.V., Podeshvo, I.V. *et al.* (2011) New polyamides with main-chain cyanine chromophores. *Polym. Sci. Ser. A Polym. Phys.*, **53**, 457–468.

91 Carothers, J.M. (2013) Design-driven, multi-use research agendas to enable applied synthetic biology for global health. *Syst. Synth. Biol.*, **7**, 79–86.

92 Shuey, S. and Shah, M. (2007) Processes for conversion of tyrosine to p-hydroxystyrene and p-acetoxystyrene. WO Patent 2,007,103,478 A2.

93 Qi, W.W., Vannelli, T., Breinig, S., Ben-Bassat, A. *et al.* (2007) Functional expression of prokaryotic and eukaryotic genes in *Escherichia coli* for conversion of glucose to p-hydroxystyrene. *Metab. Eng.*, **9**, 268–276.

94 Bor-Sen, C., Wu, W.-S., Wang, Y.-C., and Wen-Hsiung, L. (2007) On the robust circuit design schemes of biochemical networks: steady-state approach. *IEEE Trans. Biomed. Circuits Syst.*, **1**, 91–104.

95 Chubukov, V., Zuleta, I.A., and Li, H. (2012) Regulatory architecture determines optimal regulation of gene expression in metabolic pathways. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 5127–5132.

96 Pan, T. and Sosnick, T. (2006) RNA folding during transcription. *Annu. Rev. Biophys. Biomol. Struct.*, **35**, 161–175.

97 Arraiano, C.M., Mauxion, F., Viegas, S.C., Matos, R.G. *et al.* (2013) Intracellular ribonucleases involved in transcript processing and decay: precision tools for RNA. *Biochim. Biophys. Acta*, **1829**, 491–513.

98 Saltelli, A., Ratto, M., Andres, T., Campolongo, F. *et al.* (2007) *Global Sensitivity Analysis. The Primer*, John Wiley & Sons, Ltd., Chichester.

99 Nelson, P. and Yang, S. (1988) Some properties of Kendall's partial rank correlation coefficient. *Stat. Probab. Lett.*, **6**, 147–150.

100 Paige, J.S., Nguyen-Duc, T., Song, W., and Jaffrey, S.R. (2012) Fluorescence imaging of cellular metabolites with RNA. *Science (New York, N.Y.)*, **335**, 1194.

101 Mutalik, V.K., Guimaraes, J.C., Cambray, G., Lam, C. *et al.* (2013) Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods*, **10**, 354–360.

102 Suess, B. (2003) Conditional gene expression by controlling translation with tetracycline-binding aptamers. *Nucleic Acids Res.*, **31**, 1853–1858.

103 Suess, B., Fink, B., Berens, C., Stentz, R. *et al.* (2004) A theophylline responsive riboswitch based on helix slipping controls gene expression in vivo. *Nucleic Acids Res.*, **32**, 1610–1614.

104 Desai, S.K. and Gallivan, J.P. (2004) Genetic screens and selections for small molecules based on a synthetic riboswitch that activates protein translation. *J. Am. Chem. Soc.*, **126**, 13247–13254.

105 Ogawa, A. and Maeda, M. (2008) An artificial aptazyme-based riboswitch and its cascading system in *E. coli*. *ChemBioChem*, **9**, 206–209.

106 Callura, J.M., Dwyer, D.J., Isaacs, F.J., Cantor, C.R. *et al.* (2010) Tracking, tuning, and terminating microbial physiology using synthetic riboregulators. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 15898–15903.

107 Callura, J.M., Cantor, C.R., and Collins, J.J. (2012) Genetic switchboard for synthetic biology applications. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 5850–5855.

108 Rodrigo, G., Landrain, T.E., and Jaramillo, A. (2012) De novo automated design of small RNA circuits for engineering synthetic riboregulation in living cells. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 15271–15276.

109 Komatsu, Y., Yamashita, S., Kazama, N., Nobuoka, K. *et al.* (2000) Construction of new ribozymes requiring short regulator oligonucleotides as a cofactor. *J. Mol. Biol.*, **299**, 1231–1243.

110 Burke, D.H., Ozerova, N.D.S., and Nilsen-Hamilton, M. (2002) Allosteric hammerhead ribozyme TRAPs. *Biochemistry*, **41**, 6588–6594.

111 Penchovsky, R. and Breaker, R.R. (2005) Computational design and experimental validation of oligonucleotide-sensing allosteric ribozymes. *Nat. Biotechnol.*, **23**, 1424–1433.

112 Klauser, B. and Hartig, J.S. (2013) An engineered small RNA-mediated genetic switch based on a ribozyme expression platform. *Nucleic Acids Res.*, 1–11.

113 Elgart, V., Jia, T., and Kulkarni, R. (2010) Quantifying mRNA synthesis and decay rates using small RNAs. *Biophys. J.*, **98**, 2780–2784.

114 Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.

115 Markham, N.R. and Zuker, M. (2008) UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol. (Clifton, N.J.)*, **453**, 3–31.

116 Zuker, M. and Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.

117 Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H. *et al.* (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.

118 Proctor, J.R. and Meyer, I.M. (2013) COFOLD: an RNA secondary structure prediction method that takes co-transcriptional folding into account. *Nucleic Acids Res.*, **41**, e102.

119 Xayaphoummine, A., Bucher, T., and Isambert, H. (2005) Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Res.*, **33**, W605–W610.

120 Geis, M., Flamm, C., Wolfinger, M.T., Tanzer, A. *et al.* (2008) Folding kinetics of large RNAs. *J. Mol. Biol.*, **379**, 160–173.

121 Lucks, J.B., Mortimer, S.A., Trapnell, C., Luo, S. *et al.* (2011) Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 11063–11068.

122 Kertesz, M., Wan, Y., Mazor, E., Rinn, J.L. *et al.* (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**, 103–107.

123 Mortimer, S.A. and Weeks, K.M. (2007) A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J. Am. Chem. Soc.*, **129**, 4144–4145.

124 Hajdin, C.E., Bellaousov, S., Huggins, W., Leonard, C.W. *et al.* (2013) Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 5498–5503.

125 Aviran, S., Trapnell, C., Lucks, J.B., Mortimer, S.A. *et al.* (2011) Modeling and automation of sequencing-based characterization of RNA structure. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 11069–11074.

126 Overgaard, M., Johansen, J., Møller-Jensen, J., and Valentin-Hansen, P. (2009) Switching off small RNA regulation with trap-mRNA. *Mol. Microbiol.*, **73**, 790–800.

127 Figueroa-Bossi, N., Valentini, M., Malleret, L., Fiorini, F. *et al.* (2009) Caught at its own game: regulatory small RNA inactivated by an inducible transcript mimicking its target. *Genes Dev.*, **23**, 2004–2015.

128 Small, I.D. and Peeters, N. (2000) The PPR motif – a TPR-related motif prevalent in plant organellar proteins. *Trends Biochem. Sci*, **25**, 46–47.

129 Boch, J., Scholze, H., Schornack, S., Landgraf, A. *et al.* (2009) Breaking the code of DNA binding specificity of TAL-type III effectors. *Science (New York, N.Y.)*, **326**, 1509–1512.

130 Barkan, A., Rojas, M., Fujii, S., Yap, A. *et al.* (2012) A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins. *PLos Genet.*, **8**, e1002910.

131 Pfalz, J., Bayraktar, O.A., Prikryl, J., and Barkan, A. (2009) Site-specific binding of a PPR protein defines and stabilizes 5′ and 3′ mRNA termini in chloroplasts. *EMBO J.*, **28**, 2042–2052.

132 Prikryl, J., Rojas, M., Schuster, G., and Barkan, A. (2011) Mechanism of RNA stabilization and translational activation by a pentatricopeptide repeat protein. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 415–420.

133 Mohanty, B.K. and Kushner, S.R. (2006) The majority of *Escherichia coli* mRNAs undergo post-transcriptional modification in exponentially growing cells. *Nucleic Acids Res.*, **34**, 5695–5704.

134 Bricker, A.L. and Belasco, J.G. (1999) Importance of a 5′ stem-loop for longevity of papA mRNA in *Escherichia coli*. *J. Bacteriol.*, **181**, 3587–3590.

11

Small Functional Peptides and Their Application in Superfunctionalizing Proteins

Sonja Billerbeck

Columbia University, Department of Chemistry, 550 West 120th Street, New York, NY 10027, USA

11.1 Introduction

Small peptides, with as few as 6 aa's (amino acids), can already bear a variety of functionalities for large areas of application. A prime example is peptides that bind fluorophores for color labeling of proteins. Established over 20 years ago, the fusion to fluorescent proteins revolutionized our ability to equip a protein with optical traits to follow its behavior *in vivo* [1]. Although fluorescent proteins persist as indispensable tools for *in vivo* imaging, their large size can in certain cases interfere with a protein's function. Small peptide tags, which directly bind fluorophores, have therefore shifted into focus and have proven to be of decisive importance for visualizing and characterizing biological systems and processes *in vivo* and *in vitro* [2]. Besides imaging, other important applications of small functional peptides include affinity tags for protein purification and interaction studies or peptides that serve as substrates for proteolytic activities, enabling the control over protein turnover for synthetic biology applications [3–5].

Besides such applications more common in basic research, small peptides have proven important as pharmaceutically relevant agents. Small antimicrobial peptides show potential as novel broad-range antibiotics and are already used in food industry, while the immune modulatory effects of peptides in humans point toward applications as new immune therapeutic agents [6]. Peptide epitopes or carbohydrate-mimicking peptides (CMPs), when inserted into a protective protein scaffold, show potential to turn into novel vaccines [7, 8]. To this end, not the modified protein is central but rather the peptide itself with the protein being used as a mere scaffold.

Traditionally, functional peptides are simply stitched onto either the N- or C-terminus of a protein. However, it is often essential to insert the peptide into the middle of the protein at a permissive site, which accepts additional aa's. Reasons for this include the situation where (i) termini of a protein might be functionally relevant or not accessible [9–11]; (ii) internal fusions might be

Synthetic Biology: Parts, Devices and Applications, First Edition. Edited by Christina Smolke.

more resistant to proteolytic degradation than terminal fusions [12]; (iii) the peptide might need to be structurally stabilized to exhibit its function as it is especially the case in vaccine development using epitope tags or small molecule mimicking peptides [13–16]; or (iv) the specific function delivered by the peptide needs *per se* to be introduced into the middle of the protein as it is the case for the engineering of cleavable proteins by insertion of a protease cleavage site [5, 9, 17, 18].

For these reasons, the focus of the subsequent discussion will be on proteins where the peptides were integrated into the sequence rather than positioned at the N- or C-terminal end.

11.2 Permissive Sites and Their Identification in a Protein

Sites within proteins at which large insertions are tolerated without loss of structural integrity and activity represent an extreme in the spectrum of sequence flexibility and have been called permissive sites [19]. Although it was originally assumed that permissive sites generally correspond to surface regions at which the added sequences do not disrupt overall folding [20], it is in fact hard to predict rationally where insertions will be tolerated even in the presence of detailed structural knowledge. It is also not clear whether all proteins and enzymes are similarly tolerant to insertions. Traditional ways to explore the structural flexibility of a protein and to identify permissive sites have therefore been random library approaches. A few studies suggest that permissive sites can be identified more rationally through comparative sequence analysis [11, 21, 22].

Two main library approaches have been applied for the identification of permissive sites in various proteins: the first is to generate insertions by limited digestion of a plasmid-encoded target gene using different restriction enzymes and religating it with a resistance cassette to be able to select for successful insertions. The resistance cassette needs to be flanked by unique restriction sites to subsequently excise the cassette and leave the gene of interest with a defined oligonucleotide insertion. The method does not enable to completely query the possible insertion space, but depending on the number of enzymes chosen for digestion, a sufficient degree of coverage can be reached [23, 24]. The second approach is insertion mutagenesis mediated by transposons, also known as linker insertion mutagenesis. Transposons are mobile genetic elements, which quasi-randomly insert in any DNA sequence mediated by the action of its corresponding transposase. Like this, any sequence can be delivered, ideally randomly, into a gene of interest as long as it is placed between the two transposase-specific recognition sites. In the simplest case, the transposon consists of a resistance marker, which is flanked by unique restriction sites as well as the transposase recognition sites. Subsequent excision of the resistance marker results in a characteristic in-frame fingerprint, which is composed of sequences from the restriction sites, the transposon ends, and target site nucleotides that were duplicated during the primary transposition event [18]. In addition,

virtually any user-defined sequence can be added into the transposon design such that it remains in the gene after excision of the selection marker.

By these random approaches, permissive sites that accept short three-residue linker insertions have been explored for the enzymes β-lactamase and β-galactosidase, revealing a variety of phenotypes depending on the nature of the inserted residues [23, 25]: insertions with similar physicochemical character as the neighboring aa's (regarding hydrophobicity, acidity, and charge) had less effect on enzyme functionality than physicochemical distant residues.

In another study, linker insertion mutagenesis of TEM1 β-lactamase revealed that two residue insertions into predicted β-sheets abolished enzymatic activity, while insertions into predicted reverse turns only affected the degree of activity but did not completely cause loss of function [26]. Further, in some cases, insertions of four residues abolished enzymatic activity, while insertion of two residues into the same site did not cause complete loss of lactamase activity.

Permissive sites accepting longer insertions – like the seven-residue-long tobacco etch virus (TEV) protease cleavage site – and were in addition accessible for efficient cleavage by the corresponding protease were explored for a variety of integral membrane proteins (like the pullulanase secretin PulD from *Klebsiella oxytoca* [27], the protein transporter FhaC of *Bordetella pertussis* [28], and *Escherichia coli lac* permease [20]) in order to study structure–function relationships. Further, random insertions of a 31-residue mostly hydrophilic peptide – so-called i31 libraries – were studied for the membrane-inserted maltose transporters MalG and MalF [29]. In both cases – for TEV cleavage site insertions as well as for i31 insertions, permissive sites allowing functional insertions were found to be mostly located in periplasmic turns or surface loops but not in parts spanning the membrane or in regions necessary for multimerization with interaction partners. Sequence insertions into nonpermissive sites affected folding, membrane insertion, multimerization, and overall functionality.

Only in a few cases permissive sites were successfully explored for cytosolic proteins. The same i31 libraries as mentioned previously were used for the mapping of functional domains and further for the identification of permissive sites in the cytosolic adenosine triphosphate (ATP)-binding component of the maltose ATP-binding cassette (ABC) transporter of *E. coli* MalK, the regulator of the *lac* operon LacI, and the F-plasmid-derived relaxase TraI [29–32].

Further, a random transposon-based approach was used to successfully identify permissive sites in the essential *E. coli* chaperonin GroEL by delivering a TEV cleavage site through transposon mutagenesis [9] as well as in the essential *Saccharomyces cerevisiae* glycosylphosphatidylinositol (GPI)-anchored membrane protein Dcw1 [10].

Identification of permissive sites within the mentioned proteins, all of which are spatially rather complex assemblies, indicates that other less challenging proteins might be able to accept even larger insertions at certain positions. However, as it was already shown for short insertions [23], the permissiveness of a certain site depended on the size and the character of the inserted sequence and the functionality of a certain insertion needed to be evaluated for each case.

Still, the current literature suggests the widespread existence of permissive sites for peptides of a length between a few and a few dozen residues.

11.3 Functional Peptides

11.3.1 Functional Peptides that Act as Binders

Peptides can specifically bind to small molecules, metals, or proteins. The ability of peptides to bind small molecules is often exploited for optical purposes. One of the most widely applied peptide tags and the first described alternative to the color labeling by protein fusions with fluorescent proteins was the small 6 residue-long tetracysteine tag (TC-tag) with the sequence CCPGCC. The TC-tag was rationally designed to covalently bind the arsenic green fluorescent dye FlAsH (fluorescein arsenical helix binder) whose fluorescence increases 1000-fold upon binding to the polypeptide tag. By now, a number of different bisarsenical fluorophores and corresponding tags have been developed [33–35]. Redesign of the FlAsH binding motif CCPGCC to bind the cyan dye AsCy3 furthermore allows for simultaneous multiple-color labeling [35]. The AsCy3 binding motif has the sequence CCKAEAACC, and discrimination between the two dyes is based on the larger interatomic distance between the two arsenics in AsCy3 (14.5 Å) than in FlAsH (6 Å). Due to its small size, the TC-tag and its derivatives have already resulted useful as a tool for *in vivo* imaging in bacterial [36] as well as eukaryotic cells [37], enabling experiments not possible with large fluorescent protein reporters. However, the method suffers from high background labeling by binding of the arsenic dyes to thiol-rich biomolecules, and extensive washing steps need to be applied to gain highly specific labeling [38].

Besides the TC-tag, other fluorophore-binding peptides have been developed: generally known as affinity tags for protein purification, the 6x histidine tag (6xHis-tag) was shown to bind metal–nitrilotriacetate (NTA)–chromophore conjugates [39] and a zinc-chelating membrane-impermeable fluorophore called HisZiFit [40]. This enabled the site-specific labeling and tracking of the stromal interaction molecule STIM1, a membrane protein for which an N-terminal fluorescent protein fusion had been shown to interfere with surface exposure [40], exemplifying again the advantage of peptide tags over protein fusions. However the binding affinity of the mentioned dyes to the 6xHis-tag was only moderate and restricted the application to extracellular labeling of cell surface proteins.

Next to these rational approaches for the design of labeling tags, directed evolution was shown to yield peptides with binding properties. Phage display was used to evolve a peptide tag that binds the dye Texas Red (called "Texas Red aptamer") and its calcium-sensing derivative X-rhod with high affinity. This way, the authors developed a 28-residue-long calcium sensor that can be "hijacked" to various cell compartments depending on the cellular localization of the protein to which the Texas Red aptamer is fused [41].

The same approach was used to evolve a lanthanide binding peptide (LBT), specifically a terbium(III)-binding peptide, of 15 residue lengths for luminescence studies [42, 43]. LBTs that bind different lanthanide ions had already been employed before for NMR studies [44] or X-ray crystallography [45]. Interestingly it was shown that the insertion of LBTs into internal loops of a protein helped in rigidifying the peptide. This made internal fusions superior to terminal fusions

for phase determination in X-ray crystallography and potentially also for the structure determination of large protein complexes by NMR [13].

Peptides can exhibit high affinity for metals, other peptides, tags, or proteins. Such affinity tags are already widely applied for protein purification, immobilization, and pull-down studies [46]. Monoclonal antibodies for most affinity tags are commercially available, making (parallel) detection of tagged proteins possible, thus circumventing the need for protein-specific antibodies. Internal tagging expands these applications to proteins with functionally relevant or inaccessible termini. An affinity tag, which is considered to be particularly suited for insertion into internal permissive sites, is the small, uncharged Strep-tag. The nine-residue peptide sequence exhibits intrinsic affinity toward Strep-Tactin, a specifically engineered streptavidin [47]. Due to the highly specific but non-covalent binding, proteins can be purified under physiological conditions in one step from crude cell lysates, without the need for high salt concentrations or other additives [48].

11.3.2 Peptide Motifs that are Recognized by Labeling Enzymes

Peptides can serve as specific substrates for enzymes. Highly specific binding of small molecules to peptides is often hard to achieve. Like this, enzyme-mediated labeling has received attention as an alternative methodology to tag cellular proteins with chemical probes. Here, enzymes selectively act on a specific peptide sequence to covalently add their cognate substrate. One such enzyme is lipoic acid ligase (LpIA) from *E. coli*. Naturally responsible for attaching lipoid acid to proteins involved in oxidative metabolism [49], LplA was rationally redesigned to specifically attach useful small molecule probes – such as alkyl azides [50] and photo-cross-linkers [51] – onto an engineered 22 aa's long LplA acceptor peptide (LAP1). The authors used this technology to label cell surface proteins and to map protein–protein interactions *in vitro.* Using yeast display for affinity selection, the originally used engineered 22-residue acceptor peptide LAP1 could be resized to only 13 residues (LAP2) while at the same time showing a 70-fold higher catalytic efficiency (k_{cat}/K_m) as a substrate for LplA [52]. By structure-guided mutagenesis, LplA was then further evolved to accept a fluorescent coumarin derivative instead of a lipoic acid derivative as substrate [53]. The resulting variant LplAW37V – together with the optimized acceptor peptide LAP2 – made the method suitable for *in vivo* labeling of proteins in eukaryotic cells. In contrast to the previously used fluorescent lipoic acid derivatives, the employed coumarin derivatives were orthogonal to eukaryotic metabolism. The original LAP1 peptide, which had been employed in cell surface and *in vitro* assays, was then revisited to develop an *in vivo* protein–protein interaction assay: the relatively low affinity but good catalytic activity of LplA for LAP1 allowed to render fluorophore attachment protein–protein interaction dependent [54]. Attachment of LAP1 and LplA to potential dimerizing partners allowed to sufficiently discriminate between an interacting and a noninteracting protein pair according to the labeling efficiency [54]. Altogether, two powerful alternative methods for the highly specific but unobtrusive labeling of proteins for imaging and interaction studies *in vivo* were introduced: PRIME (PRobe Incorporation Mediated

by Enzymes) and ID-PRIME (Interaction-Dependent PRobe Incorporation Mediated by Enzymes). Meanwhile, several improvements regarding kinetics of labeling and diversification on the derivatization of the PRIME substrates were reported [55–58] as well as examples for their application in solving biological questions [59].

A similar labeling strategy is based on the biotin ligase BirA from *E. coli* [60] or its analogs from yeast (yBL) and *Pyrococcus horikoshii* (PhBL) [61]. Biotin ligases covalently attach biotin to a lysine in a 15-residue biotin acceptor peptide (BAP) [62]. Like this, biotin analogs that are accepted by BirA can be attached to proteins labeled with BAP [63]. Although BirA action is orthogonal to eukaryotic biotinylation, the method is essentially restricted to protein labeling on the cell surface as endogenous biotin still serves as a much better substrate than the corresponding derivatives. The BirA/BAP pair has been used for proximity studies on cell surfaces and to image communication across cells by transsynaptic biotinylation [64].

Clearly each peptide-based labeling method has its advantages regarding host range, complexity of the labeling approach, and optical properties of the employed fluorescent or luminescent labels. Thus, the choice of the appropriate labeling technique needs to be carefully considered when designing an *in vivo* labeling experiment. Some excellent recent reviews elaborate in more detail on these topics [65–67]. So far, labeling of protein at internal sites has not been extensively explored, although it would add more flexibility in experimental design than only focusing on terminal tagging.

11.3.3 Peptides as Protease Cleavage Sites

Peptides can be used to influence the degradation kinetics of proteins for applications in basic science, synthetic biology, and biotechnology [4, 5, 68, 69]. Tuning a protein's stability *in vivo* can be achieved through N- or C-terminal degradation tags [70–72]. Alternatively, the process of protein inactivation can be rendered conditionally, for example, by the insertion of a protein cleavage site into an internal permissive site, which is recognized by a specific (ideally host-orthogonal) protease. Nuclear inclusion protein a (NIa) proteases obtained from viruses of the family Potyviridae are the most promising target proteases due to their high activities and sequence specificity. Potyvirus proteases are responsible for processing the potyvirus polyprotein into its functional units [73]. The best-characterized and most commonly applied member is TEV protease, which recognizes the consensus sequence ENLYFQG, with cleavage according between Q (P1 positions) and G (P1*) position. TEV is relaxed toward substitutions in the P5, P4, P2, and P1* position [74, 75]. This gives some freedom for cleavage site design, although cleavage efficiencies vary depending on the exact residue inserted.

The TEV consensus sequence is not found in the proteome of mammalian cells [76], yeast [4], or *E. coli* [77]. Besides its application in protein purification where it is frequently used to cleave off affinity tags [78], proteolysis by TEV protease was used for *in vivo* studies as a tool to bleach essential proteins [4, 77], to study phosphorylation-dependent protein–protein interactions [79], to trigger

apoptosis in mammalian cells [76], or to establish gene networks for synthetic biology [5, 17]. Further well-characterized potyvirus proteases are the plum pox virus (PPV) protease, which recognizes the 7 aa sequence NVVVHQ/A [80] and the tobacco vein mottling virus (TVMV) protease being specific for ETVRFQG/S [81]. While PPV and TVMV proteases have been used for processing of fusion proteins *in vivo* and *in vitro* [82, 83], they have not yet been extensively explored as tools for systemic functionality studies of target proteins *in vivo*. However, potyvirus proteases are orthogonal to each other, meaning they cannot recognize the cleavage sites of each other efficiently [74, 84], which might facilitate combined employment *in vivo*, for example, for synthetic posttranslational modification networks.

11.3.4 Reactive Peptides

All before mentioned peptides offer "passive" functions like "binding" or "being recognized" and are acted upon by separately encoded enzymes. Peptides, which encode an "active" function, for example, catalytic activity, constitute an interesting expansion to the functional portfolio of peptides, especially when these activities could be added in an orthogonal manner to the activity of an already functional scaffold protein.

One outstanding example of a reactive peptide is the 13 aa SpyTag peptide that rapidly forms an isopeptide bond between a peptide-internal lysine and an aspartate residue in its target protein SpyCatcher (138 aa, 15 kDa) [85, 86]. In the demonstrated setting, SpyTag was reactive irrespective of the location in the scaffold protein (terminal or internal). Recently, a second orthogonal isopeptide bond forming peptide tag/target protein pair (SnoopTag/SnoopCatcher) was introduced that is completely orthogonal to the SpyTag/SpyCatcher system [87]. Both pairs have been used together to build synthetic polyproteins [87], to design optimized vaccines [88], and to assemble bioactive protein hydrogels [89]. The hydrogel assembly was achieved by combining internal and terminal SpyTag insertions.

11.3.5 Pharmaceutically Relevant Peptides: Peptide Epitopes, Sugar Epitope Mimics, and Antimicrobial Peptides

The structural flexibility of proteins to accept additional residues makes them suitable scaffolds to structurally stabilize or protect peptides from degradation. While the functional peptides that were discussed before can be seen as tools to facilitate the study of properties or the (modulation of the) *in vivo* behavior of a certain protein of interest, the functional peptides of the following section will themselves be the actual targets of interest, and the protein in which it is inserted is a tool to facilitate its production or application. This section is only meant to give a notion on the diversity of pharmaceutically relevant functions that can be adopted by peptides and to discuss the potential impact that an extended knowledge about permissive sites could make to the field of therapeutic peptides. For a broader treatment, the reader is referred to an excellent recent review [90].

11.3.5.1 Peptide Epitopes

The immune response elicited by a given pathogen is specific against certain exposed fractions of the pathogen's proteome, called epitopes. For the design of novel vaccines, approaches are explored where known epitopes are taken out of their natural (pathogenic) context and inserted into a different protein scaffold that is, in contrast to the protein of the epitope's origin, nontoxic and easy to purify in high yield. It was already shown a decade ago that a permissive site within *B. pertussis* adenylate cyclase toxin-hemolysin (ACT-Hly) can be used to deliver a $CD8^+$ T-cell epitope into antigen-presenting cells *in vivo* and induce protective antiviral as well as therapeutic antitumor cytotoxic T-cell responses [91–93]. Adenylate cyclase toxoids can penetrate a variety of immune effector cells. Variants with disrupted catalytic activity, which are still cell invasive, are therefore considered a potent scaffold for vaccine design.

Further, the nontoxic B subunit of cholera toxin [12, 94] as well as the hepatitis B core particle-forming protein HBcAg (for hepatitis B core antigen) have been explored as potential vaccine scaffolds by insertion of relevant epitopes, for example, a hepatitis C-specific epitope or the HIV-1-neutralizing epitope [95, 96]. More recently, adenovirus (Ad) capsid proteins embody enormous promise for the realization of diverse vaccines [97–99]. Also computational strategies have been developed to guide the design of epitope-equipped protein scaffolds for conformational stabilization and immune presentation [14–16].

For the design of novel chimeric vaccines, two points can be extracted from the body of available literature that seem to be most relevant for consideration: (i) the protein that is supposed to serve as a scaffold should be highly immunogenic by itself to elicit – next to the scaffold-specific response – a high antibody production against the inserted epitope and (ii) a solvent-exposed permissive sites within the scaffold should be known. The second point seems relevant for two reasons: Firstly, though it was shown that terminal epitope fusions are in principle able to elicit epitope-specific immune responses, proteins were prone to degradation that might interfere with the generation of the response. In contrast, proteins with internal insertions were stably expressed [12]. Further and more importantly, it was shown for HBcAg-epitope chimeras that insertions in an internal permissive site showed higher epitope-specific antibody production than terminal fusions, especially when the chimeric protein was designed such that the inserted epitope replaced another immunogenic region of the scaffold [100].

Although some proteins with known permissive sites are available and employed for the design of vaccine chimeras, the field seems limited by novel scaffolds. Better knowledge on permissive sites and their identification in potentially attractive scaffolds could thus pave the way for novel vaccine strategies.

11.3.5.2 Peptide Mimotopes

Peptide mimotopes are short aa sequences that *mimic* small molecules or carbohydrates. By using a protein that naturally binds the target molecule, for example, monosaccharide-binding lectins, mimotopes can be selected from peptide libraries by phage display. Per definition, a selected peptide is considered

mimetic if it also interacts with several other proteins or receptors that are known to bind the natural ligand [101]. Mimicry is thus defined as "binding to the same proteins as the natural ligand" rather than resembling the physicochemical properties or the molecular recognition characteristics of the ligand. As mimicry is rarely obvious upon comparing the chemical structures of ligand and mimotope, rational design of a peptide small molecule mimic is currently nearly impossible.

Especially CMPs are of pharmaceutical relevance. Carbohydrates are often displayed on the outer surfaces of pathogens and tumor cells and are therefore potential immunological targets for diagnosis, antibody production, and vaccine development. However, carbohydrates are intrinsically T-cell-independent antigens, which diminish their efficacy as immunogens [102]. Further, carbohydrates are difficult to chemically synthesize in high yield particularly due to the absolute requirement for the correct stereoconfiguration [103]. In contrast, preparative production routes to peptides have emerged over the last years [104], and peptides have an absolute requirement for T cells, making them better immunogens. The conversion of carbohydrate epitopes to peptide mimotopes has therefore potential to overcome the shortcomings of carbohydrate immunogens [6].

The first attempt to establish a CMP using phage display was done with the jack bean lectin concanavalin A (ConA), which binds α-mannose. This effort led to the identification of the tripeptide YPY to which ConA binds with high affinity [105]. These studies were followed by the screening and successful identification of a variety of peptide mimics against various pathogen- and virus-associated mono- and polysaccharides of high complexity. For a detailed and comprehensive overview on available mimotopes and their applications, see [101].

The available studies support the remarkably high potential of peptides to mimic virtually any desired chemical monomer or polymer – the right peptide sequence simply needs to be discovered by an appropriate method such as phage display. Although CMPs are traditionally chemically synthesized and – to the best knowledge of this author – have not been inserted into a permissive site of a protein scaffold, they illustrate the great versatility of potential chemical characteristics and functions that peptides can adopt and that can even be expanded by directed evolution. Consequently, as exemplified for peptide epitopes, the insertion of mimotopes into protein scaffolds for structural stabilization or to simply use theses scaffolds as carrier for *in vivo* delivery or high-yield production might bear a great but unexplored potential.

11.3.5.3 Antimicrobial Peptides

Antimicrobial peptides are short, mostly cationic hydrophobic peptides with antimicrobial activity against a broad variety of microbes [106]. Although even di- and tripeptides with antimicrobial activity have been reported [107], their size usually varies from 7 [108] to about 60 aa residues [109]. Antimicrobial peptides adopt secondary structures including α-helices, relaxed coils, antiparallel β-sheets, and gamma-core motifs – two antiparallel ββ-sheets connected by a short turn as found in defensin-like peptides and often including disulfide bridges. There is a relationship between structure and function, with amphipathic

α-helical peptides being often more active than structurally less defined peptides and peptides with the gamma-core motif often being very active [110].

Due to their overall positive charge, antimicrobial peptides can accumulate at the negatively charged microbial cell surface – which often contains acidic polymers in Gram-negative as well as Gram-positive bacteria. After self-mediated uptake, they insert into the cytoplasmic membrane, disrupting its physical integrity [111]. Some peptides can also cross the membrane and act on intracellular targets [107]. Besides their high potential as broad-range antibiotics, recent studies point toward a second function: the cationic peptides – which are not only produced by bacteria in their fight to populate ecological niches but are also found in higher organisms as defense mechanism [112] – are modulators of innate immunity [113, 114]. This property is discussed to have potential for the development of novel anti-infective therapeutic strategies [115]. A comprehensive database containing the sequences and properties of animal and plant peptides is available [116].

To meet the needs of basic research and clinical trials, large quantities of highly purified peptides are required. Although some peptide antibiotics are synthesized non-ribosomally by complex peptide synthetases, most of the peptides are genetically encoded. Recombinant production in bacteria offers an attractive approach for cost-effective large-scale peptide manufacture. A database housing information on recombinant approaches to generate suitable amount of antimicrobial peptides for biological and structural studies has been established [117]. The field of antimicrobial peptide production therefore nicely exemplifies the attempt to overcome shortcomings in the chemical production of peptides – a field that could also be of great interest for the production of mimotopes.

Most production approaches resemble the natural production mechanism of antimicrobial peptides in their host [118]: to protect the production host from peptide toxicity and the peptide from cellular degradation, the target peptides are produced as larger precursors that are then processed by proteases to release the actual active peptide moiety. Like this a variety of fusion partners and release strategies have been explored [119]. Besides host toxicity and degradation, also the intrinsic hydrophobicity of the peptides impairs its soluble production when overproduced in a bacterial host. Commonly used strategies involve solubility-enhancing fusion partners like thioredoxin and glutathione-*S*-transferase (GST) [120], but also small aggregation-promoting carriers, for example, PurF [121] or ketosteroid isomerase [122], were explored. The rationale for the latter is to channel the peptide fusion into inclusion bodies to circumvent host toxicity and degradation while still having easy access to the peptide. Release from a carrier can be achieved by chemical hydrolysis or by specific proteases like TEV protease [119].

However, current yields for purified peptides are limited to milligrams per liter culture and current efforts focus on finding novel scaffolds for efficient expression. Again to the best knowledge of this author, peptide production using permissive site within solubilizing scaffold has not been explored yet, but seems to be a promising alternative to current approaches, especially to address the problem of peptide degradation during production.

11.4 Conclusions

Literature suggests an astonishing versatility in peptide functionalities (Table 11.1). Now standard directed evolution and selection techniques have the potential to amplify this spectrum. Indeed, successful peptide engineering by directed evolution has already been achieved as exemplified by the identification of novel chromophore-binding peptides or peptide mimotopes from phage libraries.

Still, the exploitation of the full potential of functional peptides for the engineering of synthetic chimeras seems to be limited by the available knowledge on permissive sites and the need for relatively labor-intensive methods to identify them in a scaffold of interest. Therefore more comprehensive rational methods would be desirable that, together with recent advances in DNA modification methods on chromosome-level [123–125], might be a step toward the exploitation of the full potential of superfunctionalized proteins.

Table 11.1 Available functional peptide tags.

Tag	Function substrate/*enzyme*	Length (aa's)	Application	References
Peptide binds small molecule				
Tetracysteine tag (TC-tag)	FlAsH, ReAsH	6	Intracellular fluorescent labeling of proteins *in vivo* and *in vitro*, eukaryotic cells, and bacteria	[33]
6xHis-tag	Ni-NTA derivatives, HisZiFit	6	Extracellular labeling of proteins	[39, 40]
Texas Red aptamer	Texas Red and derivatives	38	Intracellular calcium sensing	[41]
Lanthanide binding tag	Lanthanides	15–20	Extracellular labeling and *in vitro* structural studies by NMR or X-ray crystallography	[43]
Peptide acts as recognition site for labeling proteins				
Biotin acceptor peptide (BAP)	Biotin derivatives/*biotin ligase*	22	Extra- and intercellular labeling of proteins *in vivo* and eukaryotic cells	[63]
Lipoic acid acceptor peptide (LAP1 and LAP2)	Lipoic acid derivatives, coumarin derivatives/*lipoic acid ligase*	13–22	Extra- and intracellular labeling of proteins *in vivo* and eukaryotic cells	[53, 54]
Peptide is recognized by hydrolyzing enzyme				
TEV protease recognition peptide	*TEV protease*	7	Cleavage of fusion proteins; has been explored as tool for mediated posttranslational modification in systems biology	[75]

(*Continued*)

Table 11.1 (Continued)

Tag	Function substrate/*enzyme*	Length (aa's)	Application	References
PPV protease recognition peptide	*PPV protease*	7	Cleavage of fusion proteins; has potential as TEV-orthogonal tool for *in vivo* studies	[82]
TVMV protease recognition peptide	*TVMV protease*	7	Cleavage of fusion proteins; has potential as TEV-orthogonal tool for *in vivo* studies	[83]
Peptide is an affinity tag				
Strep-tag	Strep-Tactin, avidin	8	One-step purification, immobilization, detection	[47]
Peptide is reactive				
SpyTag	SpyCatcher	13	Isopeptide bond formation with protein partner SpyCatcher	[85]
SnoopTag	SnoopCatcher	12	Isopeptide bond formation with protein partner SnoopCatcher. The reaction is orthogonal to SpyTag/SpyCatcher	[87]

Definitions

Permissive site Sites within a protein where insertions of several amino acids are accepted without compromising folding or function

Functional peptide Small amino acid sequence (defined here as approximately 6–30 residues), which shows a stand-alone biological function

Superfunctionalization Incorporation of an (orthogonal) functionality into the primary function of a protein by insertion of a functional peptide sequence into a permissive site of the target protein

Protein scaffold A protein who's primary function is to structurally serve as docking point for additional functions

Protein engineering The design of new enzymes or proteins with new or desirable functions

Abbreviations

aa	amino acid
FlAsH	fluorescein arsenical helix binder
TC-tag	tetra cysteine tag
LplA	lipoic acid ligase
LAP	LplA acceptor peptide
BirA	biotin ligase

BAP	biotin acceptor peptide
LBT	lanthanide binding peptide
TEV protease	tobacco etch virus protease
PPV protease	plum pox virus protease
TVMV	tobacco vein mottling virus
CMP	carbohydrate-mimicking peptide
GST	glutathione-*S*-transferase

Acknowledgment

I would like to thank Sven Panke for helpful discussion. The author was supported by a Junior Fellow award from the Simons Foundation.

References

1 Rizzuto, R. *et al.* (1995) Chimeric green fluorescent protein as a tool for visualizing subcellular organelles in living cells. *Curr. Biol.*, **5** (6), 635–642.
2 Lotze, J. *et al.* (2016) Peptide-tags for site-specific protein labelling in vitro and in vivo. *Mol. Biosyst.*, **12** (6), 1731–1745.
3 Pina, A.S., Batalha, I.L., and Roque, A.C. (2014) Affinity tags in protein purification and peptide enrichment: an overview. *Methods Mol. Biol.*, **1129**, 147–168.
4 Taxis, C. *et al.* (2009) Efficient protein depletion by genetically controlled deprotection of a dormant N-degron. *Mol. Syst. Biol.*, **5**, 267.
5 Calles, B. and de Lorenzo, V. (2013) Expanding the Boolean logic of the prokaryotic transcription factor XylR by functionalization of permissive sites with a protease-target sequence. *ACS Synth. Biol.*, **2** (10), 594–603.
6 Pashov, A., Monzavi-Karbassi, B., and Kieber-Emmons, T. (2009) Immune surveillance and immunotherapy: lessons from carbohydrate mimotopes. *Vaccine*, **27** (25–26), 3405–3415.
7 Guenaga, J. *et al.* (2011) Heterologous epitope-scaffold prime: boosting immuno-focuses B cell responses to the HIV-1 gp41 2F5 neutralization determinant. *PLoS One*, **6** (1), e16074.
8 Sun, Z.W. *et al.* (2016) An immunogen containing four tandem 10E8 epitope repeats with exposed key residues induces antibodies that neutralize HIV-1 and activates an ADCC reporter gene. *Emerg. Microbes Infect.*, **5** (6), e65.
9 Billerbeck, S. *et al.* (2013) Towards functional orthogonalisation of protein complexes: individualisation of GroEL monomers leads to distinct quasihomogeneous single rings. *ChemBioChem*, **14** (17), 2310–2321.
10 Zordan, R.E. *et al.* (2015) Avoiding the ends: internal epitope tagging of proteins using transposon Tn7. *Genetics*, **200** (1), 47–58.
11 Sturgill, T.W. *et al.* (2008) TOR1 and TOR2 have distinct locations in live cells. *Eukaryot. Cell*, **7** (10), 1819–1830.
12 Backstrom, M. *et al.* (1994) Insertion of a HIV-1-neutralizing epitope in a surface-exposed internal region of the cholera toxin B-subunit. *Gene*, **149** (2), 211–217.

13 Barthelmes, K. *et al.* (2011) Engineering encodable lanthanide-binding tags into loop regions of proteins. *J. Am. Chem. Soc.*, **133** (4), 808–819.

14 Correia, B.E. *et al.* (2010) Computational design of epitope-scaffolds allows induction of antibodies specific for a poorly immunogenic HIV vaccine epitope. *Structure*, **18** (9), 1116–1126.

15 Ofek, G. *et al.* (2010) Elicitation of structure-specific antibodies by epitope scaffolds. *Proc. Natl. Acad. Sci. U.S.A.*, **107** (42), 17880–17887.

16 Correia, B.E. *et al.* (2014) Proof of principle for epitope-focused vaccine design. *Nature*, **507** (7491), 201–206.

17 Copeland, M.F. *et al.* (2016) A transcription activator-like effector (TALE) induction system mediated by proteolysis. *Nat. Chem. Biol.*, **12** (4), 254–260.

18 Reznikoff, W.S. (2006) Tn5 transposition: a molecular tool for studying protein structure-function. *Biochem. Soc. Trans.*, **34** (Pt 2), 320–323.

19 Hofnung, M., Bedouelle, H., Boulain, J., Clement, J., Charbit, A. *et al.* (1988) Genetic approaches to the study and use of proteins: random point mutations and random linker insertions. *Bull. Inst. Pasteur*, **86**, 95–101.

20 Manoil, C. and Bailey, J. (1997) A simple screen for permissive sites in proteins: analysis of Escherichia coli lac permease. *J. Mol. Biol.*, **267** (2), 250–263.

21 Burg, L. *et al.* (2016) Internal epitope tagging informed by relative lack of sequence conservation. *Sci. Rep.*, **6**, 36986.

22 Schlehuber, L.D. and Rose, J.K. (2004) Prediction and identification of a permissive epitope insertion site in the vesicular stomatitis virus glycoprotein. *J. Virol.*, **78** (10), 5079–5087.

23 Barany, F. (1985) Two-codon insertion mutagenesis of plasmid genes by using single-stranded hexameric oligonucleotides. *Proc. Natl. Acad. Sci. U.S.A.*, **82** (12), 4202–4206.

24 Perlman, D. and Halvorson, H.O. (1986) The MURFI linker for multiple reading frame insertion of a sense or nonsense codon into DNA. *Nucleic Acids Res.*, **14** (5), 2139–2155.

25 Breul, A. *et al.* (1991) Linker mutagenesis in the lacZ gene of *Escherichia coli* yields variants of active beta-galactosidase. *Eur. J. Biochem.*, **195** (1), 191–194.

26 Hallet, B., Sherratt, D.J., and Hayes, F. (1997) Pentapeptide scanning mutagenesis: random insertion of a variable five amino acid cassette in a target protein. *Nucleic Acids Res.*, **25** (9), 1866–1867.

27 Guilvout, I. *et al.* (1999) Genetic dissection of the outer membrane secretin PulD: are there distinct domains for multimerization and secretion specificity? *J. Bacteriol.*, **181** (23), 7212–7220.

28 Guedin, S. *et al.* (2000) Novel topological features of FhaC, the outer membrane transporter involved in the secretion of the *Bordetella pertussis* filamentous hemagglutinin. *J. Biol. Chem.*, **275** (39), 30202–30210.

29 Nelson, B.D., Manoil, C., and Traxler, B. (1997) Insertion mutagenesis of the lac repressor and its implications for structure-function analysis. *J. Bacteriol.*, **179** (11), 3721–3728.

30 Lippincott, J. and Traxler, B. (1997) MalFGK complex assembly and transport and regulatory characteristics of MalK insertion mutants. *J. Bacteriol.*, **179** (4), 1337–1343.

31 Haft, R.J. *et al.* (2006) General mutagenesis of F plasmid TraI reveals its role in conjugative regulation. *J. Bacteriol.*, **188** (17), 6346–6353.
32 Ehrmann, M. *et al.* (1997) TnTIN and TnTAP: mini-transposons for site-specific proteolysis in vivo. *Proc. Natl. Acad. Sci. U.S.A.*, **94** (24), 13111–13115.
33 Adams, S.R. *et al.* (2002) New biarsenical ligands and tetracysteine motifs for protein labeling in vitro and in vivo: synthesis and biological applications. *J. Am. Chem. Soc.*, **124** (21), 6063–6076.
34 Bhunia, A.K. and Miller, S.C. (2007) Labeling tetracysteine-tagged proteins with a SplAsH of color: a modular approach to bis-arsenical fluorophores. *ChemBioChem*, **8** (14), 1642–1645.
35 Cao, H. *et al.* (2007) A red cy3-based biarsenical fluorescent probe targeted to a complementary binding peptide. *J. Am. Chem. Soc.*, **129** (28), 8672–8673.
36 Ignatova, Z. and Gierasch, L.M. (2004) Monitoring protein stability and aggregation in vivo by real-time fluorescent labeling. *Proc. Natl. Acad. Sci. U.S.A.*, **101** (2), 523–528.
37 Hoffmann, C. *et al.* (2005) A FlAsH-based FRET approach to determine G protein-coupled receptor activation in living cells. *Nat. Methods*, **2** (3), 171–176.
38 Hoffmann, C. *et al.* (2010) Fluorescent labeling of tetracysteine-tagged proteins in intact cells. *Nat. Protoc.*, **5** (10), 1666–1677.
39 Guignet, E.G., Hovius, R., and Vogel, H. (2004) Reversible site-selective labeling of membrane proteins in live cells. *Nat. Biotechnol.*, **22** (4), 440–444.
40 Hauser, C.T. and Tsien, R.Y. (2007) A hexahistidine-Zn2+–dye label reveals STIM1 surface exposure. *Proc. Natl. Acad. Sci. U.S.A.*, **104** (10), 3693–3697.
41 Marks, K.M., Rosinov, M., and Nolan, G.P. (2004) In vivo targeting of organic calcium sensors via genetically selected peptides. *Chem. Biol.*, **11** (3), 347–356.
42 Allen, K.N. and Imperiali, B. (2010) Lanthanide-tagged proteins – an illuminating partnership. *Curr. Opin. Chem. Biol.*, **14** (2), 247–254.
43 Martin, L.L. *et al.* (2005) Rapid combinatorial screening of peptide libraries for the selection of lanthanide-binding tags (LBTs). *QSAR Comb. Sci.*, **24** (10), 1149–1157.
44 Lee, L. and Sykes, B.D. (1980) Strategies for the uses of lanthanide NMR shift probes in the determination of protein structure in solutio. Application to the EF calcium binding site of carp parvalbumin. *Biophys. J.*, **32** (1), 193–210.
45 Weis, W.I. *et al.* (1991) Structure of the calcium-dependent lectin domain from a rat mannose-binding protein determined by MAD phasing. *Science*, **254** (5038), 1608–1615.
46 Waugh, D.S. (2005) Making the most of affinity tags. *Trends Biotechnol.*, **23** (6), 316–320.
47 Skerra, A. and Schmidt, T.G.M. (2000) Use of the strep-tag and streptavidin for detection and purification of recombinant proteins. *Methods Enzymol.*, **326**, 271–304.
48 Schmidt, T.G. and Skerra, A. (2007) The strep-tag system for one-step purification and high-affinity detection or capturing of proteins. *Nat. Protoc.*, **2** (6), 1528–1535.
49 Cronan, J.E., Zhao, X., and Jiang, Y. (2005) Function, attachment and synthesis of lipoic acid in *Escherichia coli*. *Adv. Microb. Physiol.*, **50**, 103–146.
50 Fernandez-Suarez, M. *et al.* (2007) Redirecting lipoic acid ligase for cell surface protein labeling with small-molecule probes. *Nat. Biotechnol.*, **25** (12), 1483–1487.

51 Baruah, H. *et al.* (2008) An engineered aryl azide ligase for site-specific mapping of protein–protein interactions through photo-cross-linking. *Angew. Chem. Int. Ed.*, **47** (37), 7018–7021.

52 Puthenveetil, S. *et al.* (2009) Yeast display evolution of a kinetically efficient 13-amino acid substrate for lipoic acid ligase. *J. Am. Chem. Soc.*, **131** (45), 16430–16438.

53 Uttamapinant, C. *et al.* (2010) A fluorophore ligase for site-specific protein labeling inside living cells. *Proc. Natl. Acad. Sci. U.S.A.*, **107** (24), 10914–10919.

54 Slavoff, S.A. *et al.* (2011) Imaging protein-protein interactions inside living cells via interaction-dependent fluorophore ligation. *J. Am. Chem. Soc.*, **133** (49), 19769–19776.

55 Liu, D.S. *et al.* (2012) Diels-Alder cycloaddition for fluorophore targeting to specific proteins inside living cells. *J. Am. Chem. Soc.*, **134** (2), 792–795.

56 Cohen, J.D., Zou, P., and Ting, A.Y. (2012) Site-specific protein modification using lipoic acid ligase and bis-aryl hydrazone formation. *ChemBioChem*, **13** (6), 888–894.

57 Uttamapinant, C. *et al.* (2012) Fast, cell-compatible click chemistry with copper-chelating azides for biomolecular labeling. *Angew. Chem. Int. Ed.*, **51** (24), 5852–5856.

58 Liu, D.S. *et al.* (2012) Quantum dot targeting with lipoic acid ligase and HaloTag for single-molecule imaging on living cells. *ACS Nano*, **6** (12), 11080–11087.

59 Rhee, H.W. *et al.* (2013) Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science*, **339** (6125), 1328–1331.

60 Cull, M.G. and Schatz, P.J. (2000) Biotinylation of proteins in vivo and in vitro using small peptide tags. *Methods Enzymol.*, **326**, 430–440.

61 Slavoff, S.A. *et al.* (2008) Expanding the substrate tolerance of biotin ligase through exploration of enzymes from diverse species. *J. Am. Chem. Soc.*, **130** (4), 1160–1162.

62 Beckett, D., Kovaleva, E., and Schatz, P.J. (1999) A minimal peptide substrate in biotin holoenzyme synthetase-catalyzed biotinylation. *Protein Sci.*, **8** (4), 921–929.

63 Chen, I. *et al.* (2005) Site-specific labeling of cell surface proteins with biophysical probes using biotin ligase. *Nat. Methods*, **2** (2), 99–104.

64 Thyagarajan, A. and Ting, A.Y. (2010) Imaging activity-dependent regulation of neurexin-neuroligin interactions using trans-synaptic enzymatic biotinylation. *Cell*, **143** (3), 456–469.

65 Jing, C. and Cornish, V.W. (2011) Chemical tags for labeling proteins inside living cells. *Acc. Chem. Res.*, **44** (9), 784–792.

66 Wombacher, R. and Cornish, V.W. (2011) Chemical tags: applications in live cell fluorescence imaging. *J. Biophotonics*, **4** (6), 391–402.

67 Fernandez-Suarez, M. and Ting, A.Y. (2008) Fluorescent probes for super-resolution imaging in living cells. *Nat. Rev. Mol. Cell Biol.*, **9** (12), 929–943.

68 Grilly, C. *et al.* (2007) A synthetic gene network for tuning protein degradation in *Saccharomyces cerevisiae*. *Mol. Syst. Biol.*, **3**, 127.

69 Neuenschwander, M. *et al.* (2007) A simple selection strategy for evolving highly efficient enzymes. *Nat. Biotechnol.*, **25** (10), 1145–1147.

70 Andersen, J.B. *et al.* (1998) New unstable variants of green fluorescent protein for studies of transient gene expression in bacteria. *Appl. Environ. Microbiol.*, **64** (6), 2240–2246.

71 Butz, M. *et al.* (2011) An N-terminal protein degradation tag enables robust selection of highly active enzymes. *Biochemistry*, **50** (40), 8594–8602.

72 Jungbluth, M., Renicke, C., and Taxis, C. (2010) Targeted protein depletion in *Saccharomyces cerevisiae* by activation of a bidirectional degron. *BMC Syst. Biol.*, **4**, 176.

73 Urcuqui-Inchima, S., Haenni, A.L., and Bernardi, F. (2001) Potyvirus proteins: a wealth of functions. *Virus Res.*, **74** (1–2), 157–175.

74 Tozser, J. *et al.* (2005) Comparison of the substrate specificity of two potyvirus proteases. *FEBS J.*, **272** (2), 514–523.

75 Kapust, R.B. *et al.* (2002) The P1′ specificity of tobacco etch virus protease. *Biochem. Biophys. Res. Commun.*, **294** (5), 949–955.

76 Gray, D.C., Mahrus, S., and Wells, J.A. (2010) Activation of specific apoptotic caspases with an engineered small-molecule-activated protease. *Cell*, **142** (4), 637–646.

77 Henrichs, T. *et al.* (2005) Target-directed proteolysis at the ribosome. *Proc. Natl. Acad. Sci. U.S.A.*, **102** (12), 4246–4251.

78 Kapust, R.B. and Waugh, D.S. (2000) Controlled intracellular processing of fusion proteins by TEV protease. *Protein Expression Purif.*, **19** (2), 312–318.

79 Wehr, M.C. *et al.* (2008) Analysis of transient phosphorylation-dependent protein-protein interactions in living mammalian cells using split-TEV. *BMC Biotech.*, **8**, 55.

80 Garcia, J.A., Riechmann, J.L., and Lain, S. (1989) Artificial cleavage site recognized by plum Pox potyvirus protease in *Escherichia coli*. *J. Virol.*, **63** (6), 2457–2460.

81 Sun, P. *et al.* (2010) Structural determinants of tobacco vein mottling virus protease substrate specificity. *Protein Sci.*, **19** (11), 2240–2251.

82 Zheng, N. *et al.* (2008) Specific and efficient cleavage of fusion proteins by recombinant plum pox virus NIa protease. *Protein Expression Purif.*, **57** (2), 153–162.

83 Nallamsetty, S. *et al.* (2004) Efficient site-specific processing of fusion proteins by tobacco vein mottling virus protease in vivo and in vitro. *Protein Expression Purif.*, **38** (1), 108–115.

84 Garcia, J.A. and Lain, S. (1991) Proteolytic activity of plum pox virus-tobacco etch virus chimeric NIa proteases. *FEBS Lett.*, **281** (1–2), 67–72.

85 Zakeri, B. *et al.* (2012) Peptide tag forming a rapid covalent bond to a protein, through engineering a bacterial adhesin. *Proc. Natl. Acad. Sci. U.S.A.*, **109** (12), E690–E697.

86 Li, L. *et al.* (2014) Structural analysis and optimization of the covalent association between SpyCatcher and a peptide Tag. *J. Mol. Biol.*, **426** (2), 309–317.

87 Veggiani, G. *et al.* (2016) Programmable polyproteams built using twin peptide superglues. *Proc. Natl. Acad. Sci. U.S.A.*, **113** (5), 1202–1207.

88 Liu, Z. *et al.* (2014) A novel method for synthetic vaccine construction based on protein assembly. *Sci. Rep.*, **4**, 7266.

89 Sun, F. *et al.* (2014) Synthesis of bioactive protein hydrogels by genetically encoded SpyTag-SpyCatcher chemistry. *Proc. Natl. Acad. Sci. U.S.A.*, **111** (31), 11269–11274.

90 Fosgerau, K. and Hoffmann, T. (2015) Peptide therapeutics: current status and future directions. *Drug Discovery Today*, **20** (1), 122–128.

91 Fayolle, C. *et al.* (1999) Therapy of murine tumors with recombinant *Bordetella pertussis* adenylate cyclase carrying a cytotoxic T cell epitope. *J. Immunol.*, **162** (7), 4157–4162.

92 Fayolle, C. *et al.* (1996) In vivo induction of CTL responses by recombinant adenylate cyclase of *Bordetella pertussis* carrying viral CD8+ T cell epitopes. *J. Immunol.*, **156** (12), 4697–4706.

93 Saron, M.F. *et al.* (1997) Anti-viral protection conferred by recombinant adenylate cyclase toxins from *Bordetella pertussis* carrying a CD8+ T cell epitope from lymphocytic choriomeningitis virus. *Proc. Natl. Acad. Sci. U.S.A.*, **94** (7), 3314–3319.

94 Bckstrom, M. *et al.* (1995) Characterization of an internal permissive site in the cholera toxin B-subunit and insertion of epitopes from human immunodeficiency virus-1, hepatitis B virus and enterotoxigenic *Escherichia coli*. *Gene*, **165** (2), 163–171.

95 Schodel, F. *et al.* (1996) Hybrid hepatitis B virus core antigen as a vaccine carrier moiety: I. presentation of foreign epitopes. *J. Biotechnol.*, **44** (1–3), 91–96.

96 Schodel, F. *et al.* (1996) Hybrid hepatitis B virus core antigen as a vaccine carrier moiety. II. Expression in avirulent *Salmonella* spp. for mucosal immunization. *Adv. Exp. Med. Biol.*, **397**, 15–21.

97 Worgall, S. *et al.* (2005) Protection against *P. aeruginosa* with an adenovirus vector containing an OprF epitope in the capsid. *J. Clin. Invest.*, **115** (5), 1281–1289.

98 Krause, A. *et al.* (2006) Epitopes expressed in different adenovirus capsid proteins induce different levels of epitope-specific immunity. *J. Virol.*, **80** (11), 5523–5530.

99 McConnell, M.J., Danthinne, X., and Imperiale, M.J. (2006) Characterization of a permissive epitope insertion site in adenovirus hexon. *J. Virol.*, **80** (11), 5361–5370.

100 Schodel, F. *et al.* (1992) The position of heterologous epitopes inserted in hepatitis B virus core particles determines their immunogenicity. *J. Virol.*, **66** (1), 106–114.

101 Matsubara, T. (2012) Potential of peptides as inhibitors and mimotopes: selection of carbohydrate-mimetic peptides from phage display libraries. *J. Nucleic Acids*, **2012**, 740982.

102 Slovin, S.F., Keding, S.J., and Ragupathi, G. (2005) Carbohydrate vaccines as immunotherapy for cancer. *Immunol. Cell Biol.*, **83** (4), 418–428.

103 Boltje, T.J., Buskas, T., and Boons, G.J. (2009) Opportunities and challenges in synthetic oligosaccharide and glycoconjugate research. *Nat. Chem.*, **1** (8), 611–622.

104 Vlieghe, P. *et al.* (2010) Synthetic therapeutic peptides: science and market. *Drug Discovery Today*, **15** (1–2), 40–56.

105 Scott, J.K. *et al.* (1992) A family of concanavalin A-binding peptides from a hexapeptide epitope library. *Proc. Natl. Acad. Sci. U.S.A.*, **89** (12), 5398–5402.

106 Hancock, R.E.W. and Lehrer, R. (1998) Cationic peptides: a new source of antibiotics. *Trends Biotechnol.*, **16** (2), 82–88.

107 Brogden, K.A. (2005) Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria? *Nat. Rev. Microbiol.*, **3** (3), 238–250.

108 Brogden, K.A. *et al.* (1996) Isolation of an ovine pulmonary surfactant-associated anionic peptide bactericidal for *Pasteurella haemolytica*. *Proc. Natl. Acad. Sci. U.S.A.*, **93** (1), 412–416.

109 Shamova, O. *et al.* (1999) Purification and properties of proline-rich antimicrobial peptides from sheep and goat leukocytes. *Infect. Immunol.*, **67** (8), 4106–4111.

110 Powers, J.P.S. and Hancock, R.E.W. (2003) The relationship between peptide structure and antibacterial activity. *Peptides*, **24** (11), 1681–1691.

111 Christensen, B. *et al.* (1988) Channel-forming properties of cecropins and related model compounds incorporated into planar lipid membranes. *Proc. Natl. Acad. Sci. U.S.A.*, **85** (14), 5072–5076.

112 Ganz, T. (2003) Defensins: antimicrobial peptides of innate immunity. *Nat. Rev. Immunol.*, **3** (9), 710–720.

113 Zasloff, M. (2002) Antimicrobial peptides of multicellular organisms. *Nature*, **415** (6870), 389–395.

114 Brogden, K.A. *et al.* (2003) Antimicrobial peptides in animals and their role in host defences. *Int. J. Antimicrob. Agents*, **22** (5), 465–478.

115 Hancock, R.E.W. and Sahl, H.G. (2006) Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. *Nat. Biotechnol.*, **24** (12), 1551–1557.

116 Wang, G., Li, X. and Wang, Z. (2016) APD3: the antimicrobial peptide database as a tool for research and education. Nucl. Acids Res., **44**, D1087–D1093. doi: 10.1093/nar/gkv1278.

117 Vassilevski, A.A., Kozlov, S.A., and Grishin, E.V. (2008) Antimicrobial peptide precursor structures suggest effective production strategies. *Recent Pat. Inflamm. Allergy Drug Discovery*, **2** (1), 58–63.

118 Li, Y. (2011) Recombinant production of antimicrobial peptides in *Escherichia coli*: a review. *Protein Expression Purif.*, **80** (2), 260–267.

119 Li, Y. (2009) Carrier proteins for fusion expression of antimicrobial peptides in *Escherichia coli*. *Biotechnol. Appl. Biochem.*, **54** (1), 1–9.

120 Lee, J.H. *et al.* (2000) High-level expression of antimicrobial peptide mediated by a fusion partner reinforcing formation of inclusion bodies. *Biochem. Biophys. Res. Commun.*, **277** (3), 575–580.

121 Majerle, A., Kidric, J., and Jerala, R. (2000) Production of stable isotope enriched antimicrobial peptides in *Escherichia coli*: an application to the production of a 15N-enriched fragment of lactoferrin. *J. Biomol. NMR*, **18** (2), 145–151.

122 Wang, H.H. *et al.* (2009) Programming cells by multiplex genome engineering and accelerated evolution. *Nature*, **460** (7257), 894–898.

123 Ronda, C. *et al.* (2016) CRMAGE: CRISPR optimized MAGE recombineering. *Sci. Rep.*, **6**, 19452.

124 Jiang, W.Y. *et al.* (2013) RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.*, **31** (3), 233–239.

Part III

Parts and Devices Supporting Spatial Engineering

12

Metabolic Channeling Using DNA as a Scaffold

Mojca Benčina[1,2], Jerneja Mori[1,2], Rok Gaber[1,2], and Roman Jerala[1,2]

[1] *National Institute of Chemistry, Department of Biotechnology, Hajdrihova 19, SI1000 Ljubljana, Slovenia*
[2] *Centre of Excellence EN-FIST, Trg Osvobodilne fronte 13, SI1000 Ljubljana, Slovenia*

12.1 Introduction

Increasing numbers of chemicals are produced by various genetically engineered organisms. Those organisms possess biosynthetic pathways composed of enzymes that act successively on the emerging substrate, in order to produce the final product molecule. The efficiency of biosynthetic pathways is crucial for industrial processes, and various strategies for the optimization of production strains have been undertaken thus far. The most common strategies include (i) increasing the pool of available substrates and/or overexpression of the enzymes of the limiting biosynthetic steps [1–3], (ii) introducing heterologous enzymes with preferred kinetic characteristics [4], and (iii) inhibition of the non-desired branching of biosynthetic pathways [5, 6].

Although diverse, none of aforementioned approaches guarantee the optimal arrangement of the enzymes of biosynthetic pathways inside the producing strain. Even if overexpressed, the enzymes still float randomly in the cytoplasm, which results in nonoptimal metabolite flow. In living cells, biosynthetic pathway enzymes or other functional polypeptides are often brought together into multienzyme complexes through specific interactions, membrane anchoring, or organelle targeting mechanisms. This type of organization increases the local concentration of enzymes and their substrates and products and minimizes the concentration of intermediates that may be toxic or unstable or may represent substrates for branching reactions. We can view such multienzyme complexes as autonomous units, where the evolving substrate travels from one enzyme to another without dissociating into the bulk solution. Therefore, reaction intermediates cannot be used by other competing biosynthetic pathways that synthesize non-desired side products. Due to the smaller characteristic distances between the consecutive enzymes in the pathway, reactions can run more efficiently.

DNA scaffolding is an artificial approach to the design formation of multienzyme complexes and will be described here. Similarly, the RNA molecule has been used as a scaffold for biosynthetic pathway enzymes [7]. Protein scaffolding

Synthetic Biology: Parts, Devices and Applications, First Edition. Edited by Christina Smolke.

and organelle targeting more closely imitate the formation of natural enzyme complexes and were used in the first attempts to improve the biosynthetic pathway efficiency by designed substrate channeling [8–12].

Although the number and distribution of the enzymes in a multi-protein complex [8–12] (Chapter 13) could be programmed by the sequence of a polypeptide backbone, the three-dimensional arrangement of the polypeptides may be unpredictable due to the flexibility in the peptide linkers between the interaction domains (Figure 12.1a,b, Table 12.1). Designing the polypeptide backbone with the scaffold-guided protein domains is limited by the number of available protein dimerization domains. Finally, each protein interaction domain has different conditions under which it folds and forms the functional interactions.

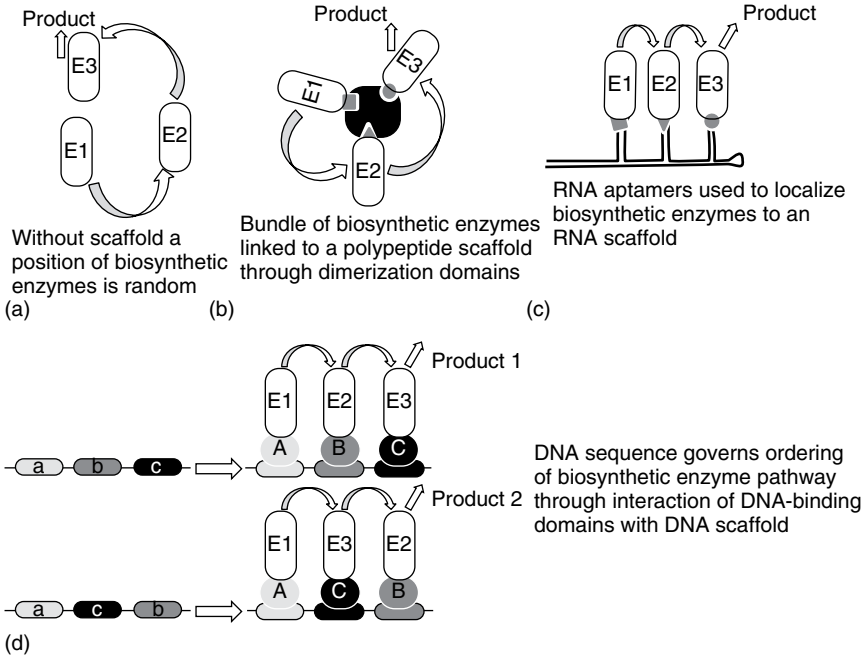


Figure 12.1 Spatial and temporal organization of biosynthetic enzymes based on different types of scaffolds. (a) Biosynthetic enzymes are typically randomly distributed inside the cell. The conversion of the substrate may therefore be limited by the diffusion rate and the concentration of the substrate and localization of the enzymes. (b) Immobilizing biosynthetic enzymes using synthetic protein scaffolds can bring the enzymes into close proximity and therefore enhance the metabolic flux. In the absence of a large superscaffold, the precise arrangement of enzymes is unpredictable and is limited by the tertiary structure of the protein scaffold. (c) Biosynthetic enzymes with predictable RNA binding sites have been assembled using synthetic RNA aptamers. The enzymes are in close proximity and in the predefined order, which enables faster conversion of the substrate into the end product. (d) An assembly line based on the DNA scaffold promotes positioning of biosynthetic enzymes in close proximity and the predefined order. The substrate conversion is faster with less unwanted side products. The enzymes are linked to DNA-binding domains, which recognize specific nucleotide sequences. The order of the enzymes can easily be changed by changing the order of the specific nucleotide sequence on the DNA program, which can lead to different end products.

Table 12.1 Advantages and disadvantages between DNA, protein, and RNA scaffolds.

Scaffold	DNA	Protein	RNA
Spatial orientation	Linear	Bundled	Linear
Order	Highly predictable	Unpredictable	Predictable, however less than for the DNA scaffold
Localization in eukaryotes	Nuclei	No limitation	Cytosol
Scaffold–enzyme ratio	Difficult to achieve substantial amount of scaffold, ratio in favor of enzymes	Easy to achieve favorable ratio with gene expression regulation	Easy to achieve favorable ratio with gene expression regulation
Scaffold–enzyme interactions	Similar, well-characterized, predictable interactions	Variations in strength, limited number of well-characterized interactions	Limited number of well-characterized RNA binding domains
Variability, number of available elements	Large number of zinc fingers and other DNA-binding domains is readily available, engineered zinc finger domains	Limited number of protein dimerization domains	Limited number of well-characterized RNA binding domains
Interference with cellular metabolism	May bind to chromatin; selecting sequences that do not affect growth	Signal transduction domains usually do not interfere in bacteria	May bind to endogenous RNA molecules; selecting sequences that do not affect growth

Niemeyer and coworkers [13] were the first to *in vitro* assemble enzymes on a DNA scaffold. They arranged NADH:FMN oxidoreductase and luciferase onto a double-stranded DNA scaffold using the biotin streptavidin linkage and showed that the immediate spatial proximity of the enzymes enhances the coupled activity. Later, they showed the operational DNA scaffold using glucose oxidase and horseradish peroxidase covalently linked to the DNA [14]. This system was further evolved by Wilner *et al.* [15], using a supramolecular DNA scaffold, who linked glucose oxidase and horseradish peroxidase via a lysine residue to the DNA oligonucleotides that hybridized onto the DNA nanostructures.

The DNA scaffold with conjugated oligonucleotides onto enzymes and assembled to DNA nanostructures is impractical to use *in vivo*. Conrado and coworkers [11] were the first to demonstrate the functional DNA scaffold in bacteria, where the enzymes were attached to the DNA-binding domains and scaffolded onto the DNA program. The principle of the DNA scaffold has some advantages in comparison with protein scaffolding (Figure 12.1c**,** Table 12.1). A DNA program sequence requires no maturation, and an ordered nucleotide binding motif can be selected at will, which provides huge orthogonality. The docking of the

anchoring of the DNA-binding protein to the DNA-target site is well characterized. Due to the close proximity of the chimeric proteins bound to the programmed nucleic acid sequence, other enzymes that might redirect synthesis are spatially excluded from the multi-protein complex. Designed DNA-binding domains can be used as fusion partners of biosynthetic pathway enzymes. These domains can share the same type of protein fold; they have similar affinity to the scaffold, and, therefore, the binding of all of the components of the biosynthetic pathway proceeds under the same reaction conditions.

RNA scaffolding is in many aspects similar to that of the DNA; however, only a limited number of well-characterized RNA binding domains is available [7, 9] (Chapter 13) (Figure 12.1d, Table 12.1). The advantage of the RNA over the DNA scaffold is that it could be used in eukaryotes to organize the metabolic pathway into the cytosol, whereas the DNA scaffold is probably limited to the prokaryotes.

Based on the pros and cons, the DNA scaffold localizes primarily in the nuclei in eukaryotic cells; therefore for eukaryotes, the protein and RNA scaffolds are the only choices. Moreover, the protein scaffold could be directed to microlocations within the cells. The main advantages of the DNA scaffold are the simple DNA program design and well-characterized anchoring of the DNA-binding proteins to the DNA-target site, as well as orthogonality; therefore, they are recommended for use in bacteria over both the RNA and protein scaffolds.

12.2 Biosynthetic Applications of DNA Scaffold

DNA scaffold-assisted biosynthesis is a viable strategy for enhancing the metabolic product yield or production rate. This enhancement appears to arise from the proximity of metabolic enzymes bound to the DNA scaffold that increases the effective concentrations of the intermediary metabolites. In every tested case, the DNA scaffold-assisted biosynthesis implemented on existing metabolic pathways improved either the product yield or rate of product synthesis (Table 12.2).

12.2.1 L-Threonine

Lee *et al*. [10] devised a DNA scaffold to facilitate the production of L-threonine in *Escherichia coli* (Figure 12.2). The biosynthetic pathway composed of the homotetramer homoserine dehydrogenase (HDH), homotetramer threonine synthase (TS), and homodimer homoserine kinase (HK) was assembled on the DNA program using 4-fingered zinc finger domains binding to 12 bp DNA-target sequences, named artificial DNA-binding domains (ADBs). Metabolic enzymes were linked to the N-terminal site of the ADBs, and they report testing several designs of the DNA program. Initially, the influence of 8, 18, and 28-bp spacers between individual DNA-target sites on the L-threonine product rate was analyzed. In addition, the impact of the target sites from one to four, for a third chimeric enzyme in the L-threonine metabolic pathway (TS-ABD3), was evaluated. The DNA scaffold with an 8 bp spacer between the DNA-target sites,

Table 12.2 DNA scaffolds used to order enzymes of biosynthetic pathways.

Product	Enzyme	DNA-binding protein and its target site	Chimeric protein	DNA scaffold[a)]			Comments	References
				Operators	No. units (n)	Spacer (bp)		
L-Threonine	Homoserine dehydrogenase (HDH)	ADB1 CAAGCTAGGGAG	HDH-ADB1 (E1)	$[1:1:1]_n$	1	8, 18, 28	Reduced production time by more than 50%	[10]
	Homoserine kinase (HK)	ADB2 GACGAGGGGGTG	HK-ADB1 (E2)	$[1:1:2]_n$ $[1:1:3]_n$ $[1:1:4]_n$	1	8		
	Threonine synthase (TS)	ADB3 GAAGGGGGGGTA	TS-ADB1 (E3)					
trans-Resveratrol	4-Coumarate–CoA ligase (4CL)	Zif268 GCGTGGGCG	Zif268-4CL (E1)	$[1:1]_n$	4, 16	2, 4, 8	More than threefold improvement in resveratrol production	[11]
	Stilbene synthase (STS)	PBSII GTGTGGAAA	PBSII-STS (E2)	$[1:1]_n$	4, 16	2, 4, 8		
1,2-Propanediol	Methylglyoxal synthase (MgsA)	ZFa GTCGATGCC	MgsA-ZFa (E1)	$[1:1:1]_n$ $[1:2:1]_n$ $[1:2:2]_n$ $[1:4:1]_n$ $[1:4:2]_n$	4, 8, 16	12	More than fourfold improvement in 1,2-propanediol production	[11]
	2,5-Diketo-D-gluconic acid reductase (DkgA)	ZFb GCGGCTGGG	DkgA-ZFb (E2)					
	Glycerol dehydrogenase (GldA)	ZFc GAGGACGGC	GldA-ZFc (E3)	$[1:1:1]_n$ $[1:2:1]_n$ $[1:2:2]_n$ $[1:4:1]_n$ $[1:4:2]_n$	1, 2, 4, 8, 16	4		

(Continued)

Table 12.2 (Continued)

Product	Enzyme	DNA-binding protein and its target site	Chimeric protein	DNA scaffold[a)]			Comments	References
				Operators	No. units (n)	Spacer (bp)		
Mevalonate	Acetoacetyl-CoA thiolase (AtoB)	ZFa GTCGATGCC	AtoB-ZFa (E1)	$[1:1:1]_n$ $[1:2:1]_n$ $[1:2:2]_n$ $[1:4:1]_n$ $[1:4:2]_n$	4, 8, 16	12	Up to threefold improvement in mevalonate production	[11]
	Hydroxymethylglutaryl-CoA synthase (HMGS)	ZFb GCGGCTGGG	HMGS-ZFb (E2)					
	Hydroxymethylglutaryl-CoA reductase (HMGR)	ZFc GAGGACGGC	HMGR-ZFc (E3)	$[1:1:1]_n$ $[1:2:1]_n$ $[1:2:2]_n$ $[1:4:1]_n$ $[1:4:2]_n$	1, 2, 4, 8, 16	4		

a) Legend: $[1:1:1]_{repeats}$ spacer(bp) [E1-spacer (bp)-E2-spacer (bp)-E3-spacer (bp)] repeated.

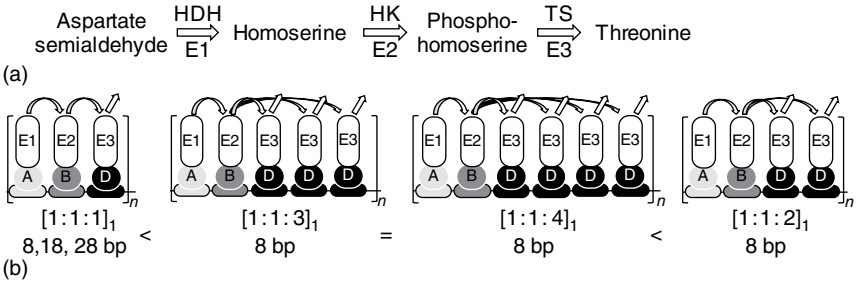


Figure 12.2 The biosynthesis of L-threonine in *E. coli* is enhanced by the DNA scaffold. (a) The three-step conversion of aspartate semialdehyde to L-threonine. (b) Arrangements of DNA-target sites on the DNA program with indicated production rates for L-threonine are depicted. The DNA scaffold includes the chimeric proteins, homoserine dehydrogenase (HDH; E1), homoserine kinase (HK; E2), and threonine synthase (TS; E3), fused to DNA-binding domains (ADB). Consecutive arrangements of DNA-target sites for threonine synthase (E3), the third enzyme in the biosynthesis of L-threonine, improved the production rate for L-threonine. The DNA-target sites specific for the individual chimeric proteins are separated with 8-, 18-, or 28-bp spacers between each DNA-target site. The fastest rate of L-threonine production in *E. coli* was obtained with the DNA program [1 : 1 : 2], with DNA-binding sites separated by 8 bp (see also [10]).

and with two copies of the TS ([1 : 1 : 2] 8 bp), reduced the production time for the L-threonine by more than 50%, with the maximum yield produced within 24 h of fermentation. For the strain without the DNA scaffold, it took 2 days to produce the same maximum yield of L-threonine. In addition, the concentration of the intermediate homoserine, which might inhibit the growth of the host cell, was reduced 15-fold.

12.2.2 *trans*-Resveratrol

We examined the ability to assemble *trans*-resveratrol (trans-3,5,4′-trihydroxystilbene) biosynthetic enzymes on DNA in the cytoplasm of *E. coli* using zinc finger DNA-binding domains, recognizing a 9-bp-long nucleotide sequence, as DNA-binding proteins [11]. The metabolic pathway for resveratrol has already been reconstituted in yeast, mammalian cells, and bacteria [7, 12, 16]. The production of the *trans*-resveratrol from 4-coumaric acid is a two-step reaction in which 4-coumaric acid is converted to 4-coumaroyl-CoA by 4-coumarate–CoA ligase (4CL). *trans*-Resveratrol is formed by the condensation of one molecule of 4-coumaroyl-CoA and three molecules of malonyl-CoA by stilbene synthase (STS) (Figure 12.3). We used a low copy number expression plasmid with genes encoding for 4CL and STS, which were fused to the C-terminus of Zif268 and PBSII zinc finger domains, respectively. The DNA scaffold was present on separate high copy number plasmids. Different DNA programs with various spacer lengths (2, 4, and 8 bp) and numbers of program repeats (4 and 16) (Table 12.2) were tested, and almost 10 $mg\,l^{-1}$ of *trans*-resveratrol was produced when the number of scaffold repeats was 4 and the spacer length between the DNA-target sites was 2 bp, which is 10 times more than with the fusion protein of 4CL and STS (Figure 12.3b) [11].

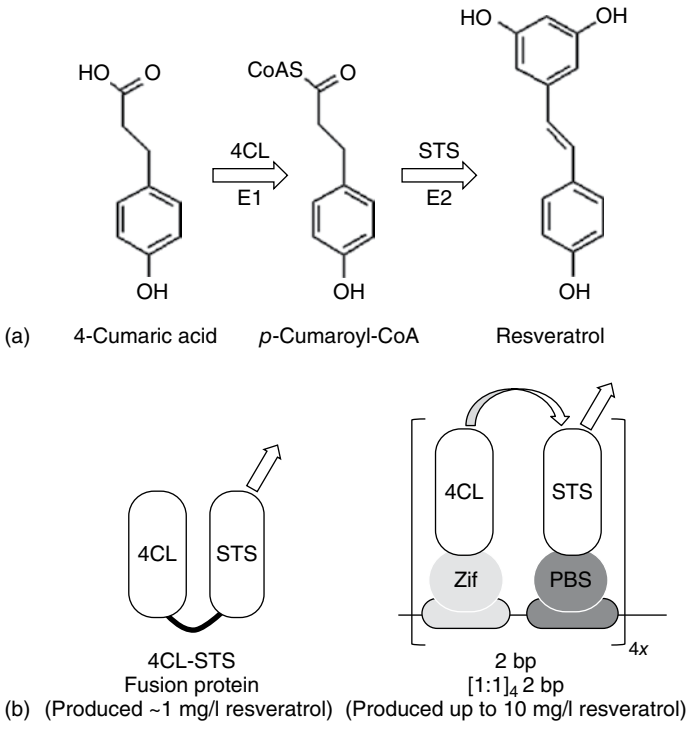


Figure 12.3 A DNA scaffold enhances the biosynthesis of *trans*-resveratrol in *E. coli*. (a) In the biosynthetic pathway of resveratrol, the 4-cumaric acid is converted to resveratrol in a two-step reaction with the biosynthetic enzymes 4-coumarate–CoA ligase (4CL) and stilbene synthase (STS). (b) Close proximity of the 4CL and STS enzymes can be achieved by fusing the enzymes with linker polypeptides or by introducing DNA scaffolds where the enzymes (4CL or STS) are fused to the DNA-binding domains (Zif268 or PBSII). The chimeric protein of the enzyme and DNA-binding domain binds to a specific nucleotide sequence present on the DNA program. The DNA-target sites specific for the individual chimeric proteins are separated with a 2-bp spacer between each of four tandem repeats [11].

12.2.3 1,2-Propanediol

A biosynthetic pathway for 1,2-propenediol composed of methylglyoxal synthase (MgsA), 2,5-diketo-D-gluconic acid reductase (DkgA), and glycerol dehydrogenase (GldA) in *E. coli* is well established [17] (Figure 12.4a). The biosynthetic enzymes were fused to the N-terminus of the zinc finger domains ZFa, ZFb, and ZFc, recognizing a 9-bp target, and the corresponding chimeras were placed on the same plasmid as the target DNA sequence [11]. Several enzyme–scaffold ratios were tested (Figure 12.4c,d), and the *E. coli* with the $[1:1:1]_4$ 12-bp spacer 1,2-propanediol system produced almost five times more product than the unscaffolded control (Table 12.2).

12.2.4 Mevalonate

The biosynthesis of mevalonate is a three-step reaction composed of acetoacetyl-CoA thiolase (AtoB), hydroxymethylglutaryl-CoA synthase (HMGS), and

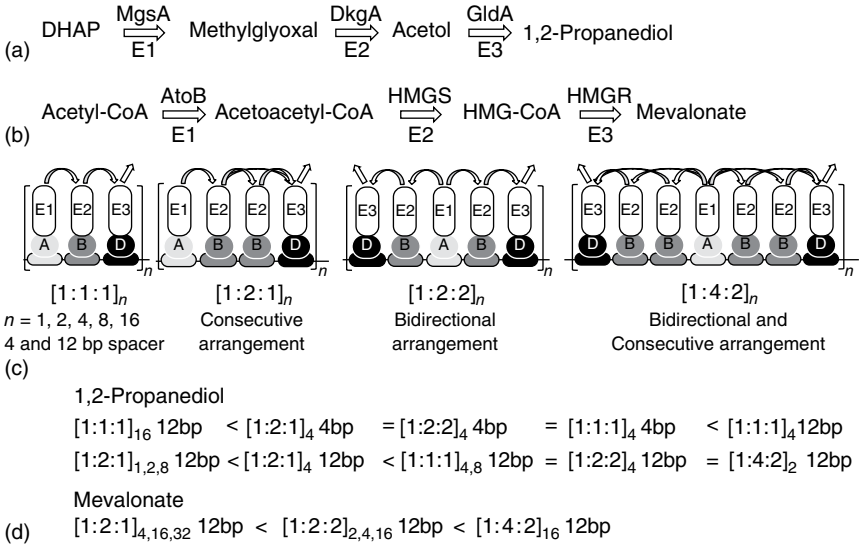


Figure 12.4 Biosynthesis of (a) 1,2-propanediol and (b) mevalonate in *E. coli*. (c) Schemes of consecutive, bidirectional, and mixed consecutive and bidirectional arrangements of DNA scaffolds with different stoichiometry and positions of DNA-target sites that were tested for the improved biosynthesis of 1,2-propanediol or mevalonate. DNA scaffolds can be used to overcome the limitations in biosynthetic pathways that occur because of individual enzymes with lower activity, compared with other enzymes in the same biosynthetic pathway. By changing the order or number of DNA-target sites, we can increase reaction yields, fine-tune biosynthesis production, and minimize side products. If the first enzyme in the biosynthetic pathway is most active, others can be distributed on both sides around the first, resulting in 1 : 2 molar ratios in favor of enzymes with low activity. Such groups of enzyme binding sites can then be multiplied on the DNA scaffold to achieve better molar ratios between the DNA scaffold and enzymes. (d) Impact of different scaffold architectures on 1,2-propanediol and mevalonate production [11].

hydroxymethylglutaryl-CoA reductase (HMGR). The biosynthesis of mevalonate in *E. coli*, as such or assisted by a protein scaffold, has already been published [8, 18]. The chimeric proteins between the enzymes of the mevalonate pathway and zinc finger domains were constructed [11] (Figure 12.4b). For the DNA scaffold design, the DNA-target sequences corresponding to each of the DNA-binding domains were placed on a separate plasmid, and the influence of the DNA-target sites arrangements on mevalonate production was tested. Similar to the resveratrol and 1,2-propanediol scaffolds, the mevalonate yield was increased up to threefold in the presence of $[1:2:2]_n$ scaffolds ($n = 2$, 4, and 16) with 12-bp spacers, compared with the random scaffold control; however, the best mevalonate yield was achieved with the DNA scaffold containing the $[1:4:2]_{16}$ program (Figure 12.4c,d).

12.3 Design of DNA-Binding Proteins and Target Sites

A self-replicating DNA plasmid in one or more copies is an ideal scaffold for any information processing; for example, the DNA sequence represents a program consisting of a series of blocks (DNA-target sites), which determine the

arrangement of the DNA-binding proteins along the DNA (Figure 12.1d). The main twist comes with the requirement that each of these DNA-binding proteins/domains is fused to a different functional protein. Therefore, the sequence of target motifs encoded by the DNA program also defines the arrangement of those functional proteins along with the order of the DNA-binding domains. Only by changing the sequence of a DNA program, either switching positions or adding new target sequences, can outcome be predicted in advance (Figure 12.1d). This requires a method for the site-specific targeting of enzymes along the DNA surface. While there are 64 nucleotide triplets in the natural code for the 20 amino acids, there could be as many as 262,144 different motifs consisting of nine nucleotides. Zinc fingers [19] and TAL elements [20] can be designed to bind to almost any desired nucleotide sequence, ranging from 9 to as many as 18 nucleotides. Additionally, we can select the target nucleotide sequence for each available DNA-binding protein.

12.3.1 Zinc Finger Domains

There are more than 700 experimentally characterized zinc fingers in the database ZIFDB [21], offering a huge selection of building elements for synthetic biology [22, 23]. Moreover, zinc fingers have similar properties, such as binding affinity or stability, which is important, since we do not need to adjust the properties of each separated part. The DNA program, therefore, represents a modular approach for various synthetic biology applications.

Up until now, only zinc finger DNA-binding domains were used to link biosynthetic proteins to DNA scaffolds. Conrado *et al*. used five different zinc finger domains (PBSII, Zif268, ZFa, ZFb, and ZFc) that were each comprised of three fingers, with a specificity for unique 9-bp DNA sequences [11, 24–26] (Table 12.2). Statistically, a 9-bp-long sequence could appear 1.2 times per genome in *E. coli*, if we assume that the nucleotide sequence distribution within the genome is random. Lee *et al*. [10] used ADBs with four fingers that recognized a 12-bp DNA sequence. All of the zinc fingers used were relatively short and bound the DNA with low nanomolar affinity. Crucially, the selected zinc finger domains should not bind functional regions of essential genes in *E. coli* or affect bacterial fitness.

As an *in vitro* test of the system components, binding to DNA can be analyzed using surface plasmon resonance (SPR) [27] (Figure 12.5a) or split GFP technology [28]. The DNA binding of the candidate zinc finger domains can be fused with split fluorescent proteins. Reassembly of the split yellow fluorescent protein (YFP) and strong fluorescence indicative of YFP reassembly occur only in the presence of a DNA scaffold that contains neighboring binding sites for, for example, PBSII and Zif268, separated by only 2 bp (Figure 12.5b) [11].

To investigate whether zinc finger domains bind their cognate DNA targets *in vivo*, a simple β-galactosidase test for DNA-binding domain activity in *E. coli* was used (Figure 12.5c). The principle of this test is that an active zinc finger domain should bind to its specific target sequence in the P_{SYN} promoter and act as a synthetic repressor, thereby decreasing the basal activity of this promoter and lowering β-galactosidase levels.

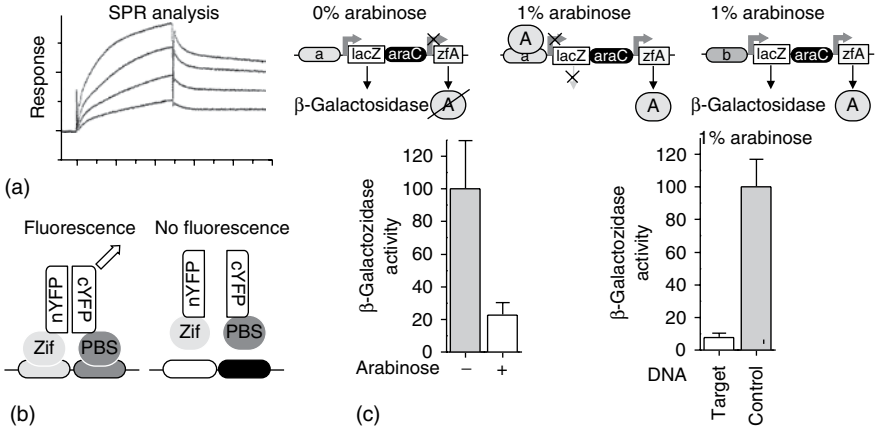


Figure 12.5 Targeting DNA *in vitro* and *in vivo* with zinc finger domains. (a) The binding affinity of zinc finger domains (e.g., Zif268) to their specific nucleotide target sequence was determined using surface plasmon resonance (SPR). With increasing concentrations of purified zinc finger protein, the response signal increases, indicating protein binding. (b) The zinc finger domain (PBSII) was fused to the N-terminal half (PBSII-nYFP), and Zif268 was fused to the C-terminal half (cYFP-Zif268) of the yellow fluorescent protein (YFP). Purified PBSII-nYFP and cYFP-Zif268 protein chimeras were mixed, either with DNA scaffolds, containing PBSII, or Zif268 target sites separated by 2-bp spacer, or DNA scaffolds with random nucleotide sequences. Fluorescence was then measured [11]. (c) The binding of the DNA-binding domain (e.g., Zif268) *in vivo* was tested with the inhibition of β-galactosidase expression. The expression of the tested zinc finger was under the control of an arabinose-inducible promoter. The *lacZ* gene was controlled by the P_{SYN} promoter, which contained either the zinc finger target site or random DNA target site (CTCTATCAATGATAGAG). β-Galactosidase activity is measured in the presence of 1% or absence (0%) of arabinose and normalized to the galactosidase levels of the unrepressed state. The β-galactosidase activity is detected when the DNA-binding protein (e.g., zinc finger A) is not expressed (no arabinose). Arabinose induces the expression of zinc finger A, which binds to the DNA-target site "a" upstream of the β-galactosidase gene, and represses the expression of β-galactosidase. If the DNA-target site "b" is not recognized by the DNA-binding protein, the expression of β-galactosidase is not affected.

Taken together, the described results indicate that (i) zinc fingers retained DNA-binding activity when fused to different proteins and (ii) two orthogonal zinc finger domains can simultaneously bind their target sequences in a DNA scaffold and bring their fused protein domains into close proximity as evidenced by the YFP reassembly.

12.3.2 TAL-DNA Binding Domains

The recent discovery of the code underlying the nucleotide sequence recognition by TAL effectors allows the design of protein domains that can bind to almost any nucleotide sequence [20] (Chapter 13). Similar to the zinc finger proteins, the TAL protein domains also seem to be ideal DNA-binding proteins for use in DNA scaffold applications. The typical TAL recognition site of 15–20 nucleotides is more than sufficient to provide the specificity required to build DNA scaffolds, even when taking into account any cross-interaction with similar

DNA-binding sequences in the host genome [29]. The binding affinity between different TAL DNA-binding domains is similar, and in the nanomolar range, as for zinc finger proteins. Due to the practically unlimited number of different combinations, there is no concern with running out of DNA-binding sites, regardless of the number of desired scaffolded enzymes.

12.3.3 Other DNA-Binding Proteins

In theory, practically any DNA-binding protein could be used in DNA scaffold applications. With the exception of zinc fingers and TALs, many of the characterized DNA-binding proteins bind DNA as dimers or tetramers (TetR, CI, and others), which would complicate the construction of DNA scaffold molecules if one desires to bind enzymes in a predefined molar ratio. Nevertheless, they may be useful for applications involving oligomeric enzymes.

12.4 DNA Program

The DNA scaffold is, in principle, more flexible in scaffold designs than protein or ssRNA scaffolds. Since dsDNA forms a helical turn approximately every 10 nt, we can use this property to guide the relative orientation of the enzymes coupled to the DNA-binding domains. We can change (i) the spatial orientation of the binding enzymes by changing the spacer length between the DNA-target sites; (ii) the number of DNA scaffold repeats, allowing us to additionally tune the biosynthetic pathway; and last but not least (iii) the DNA scaffold, which enables us to modify the enzymatic stoichiometry.

12.4.1 Spacers between DNA-Target Sites

The program DNA is designed to organize biosynthetic pathway enzymes into a functional complex. Spacers in the DNA sequence separating the DNA-target sites on the program DNA determine the spatial orientation of chimeric biosynthetic enzymes relative to each other (Figure 12.6). The binding sites for three-fingered zinc fingers span nine nucleotides but can be extended to 18-bp recognition motifs for longer zinc fingers, spanning from one to two DNA duplex helical turns, respectively. The binding sites for DNA-binding proteins are separated by spacers, which are nucleotides that are not occupied by DNA-binding proteins. The length of the spacer sequence is not coincidental, and the selection follows the three-dimensional structure of a DNA molecule. One turn of the DNA helix is 10.5 bp long, which roughly overlaps with the length of a DNA molecule encircled by one zinc finger domain recognizing and binding to 9 bp. In order to have functional units on the same side of a DNA molecule serving as a DNA program, it is of high importance to select the right spacer length. The double helix of the DNA defines on which side of the helix the functional domain will be attached, which is defined by the length of the spacer between the DNA-target sites: a spacer of one or two nucleotides positions them very close, while a

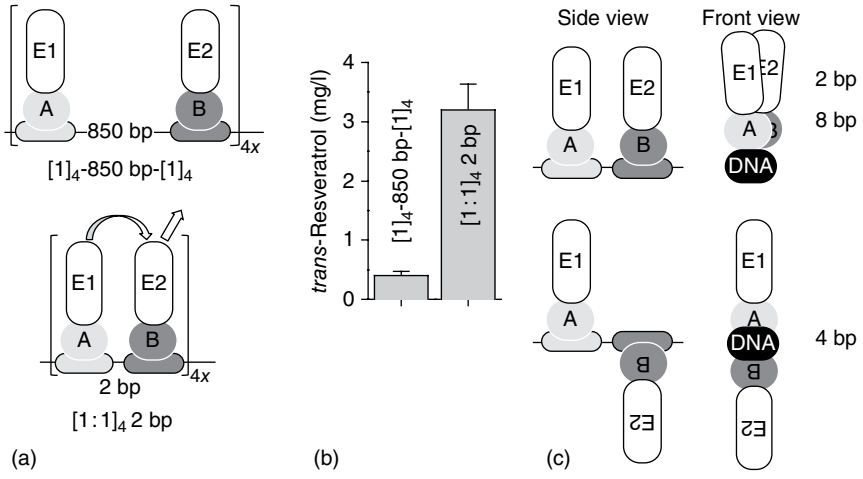


Figure 12.6 Spatial position of biosynthetic enzymes is defined by DNA program. (a) Scheme of two types of DNA scaffolds that differ in spacer lengths separating the DNA-target sites. A first scaffold plasmid (left) carries four copies of Zif268 and four copies of PBSII binding sites, separated by an insertion of 850 bp along part of the plasmid backbone. A second scaffold plasmid carries four copies of the Zif268 and PBSII binding sites separated by 2-bp-long spacers. (b) Enzyme clustering improves the production of *trans*-resveratrol, which was measured in *Escherichia coli*-expressing fusion enzymes (Zif268-4CL and PBSII-STS) with different DNA program plasmids $[1]_4$-850 bp-$[1]_4$ and $[1:1]_4$ 2-bp spacers (for details see Figure 12.4a) [11]. (c) The spatial orientation of the enzymes is governed with a spacer between the DNA-target sites. The 2- and 8-bp spacers orientate chimeric enzymes on the same side of the DNA program (up). The 4-bp spacer between the target sites orientates the enzymes on opposite sides of the DNA program (below). The best production of *trans*-resveratrol in *E. coli* was achieved when the binding sites for Zif268 and PBSII were separated with 8 bp [11].

spacer of four to five nucleotides positions the neighboring two functional domains to the opposite sides of the helix (Figure 12.6c).

Initially, the impact of the clustering of metabolic pathway enzymes [11] was examined, and the DNA-target sites within the $[1:1]_4$ scaffold were separated on the plasmid by either 2 or 850 bp (Figure 12.6a). The $[1]_4$-850 bp-$[1]_4$ scaffold provided the same number of binding sites on the plasmid for both enzymes but prevented the close proximity of the bound enzymes to one another. The fivefold enhancement in resveratrol production observed for the $[1:1]_4$ scaffold was abolished when the binding sites for each enzyme were positioned far apart on the plasmid, indicating that close proximity of the pathway enzymes is important (Figure 12.6b).

In the example of resveratrol biosynthesis, we [11] examined whether the three-dimensional positioning of individual enzymes effects production yields (Figure 12.6c). DNA scaffolds with DNA-target sites separating 2, 4, or 8 bp were constructed. Considering the standard DNA topology, the 2- and 8-bp spacer position functional units were on the same site of the DNA scaffold, and the 4-bp spacer position functional units to the opposite site of the DNA program. In the case of the $[1:1]_{16}$ resveratrol system, the best product yields were obtained with

the DNA program, in which individual DNA-target sites were separated with spacer lengths of 2 or 8 bp, while a spacer length of 4 bp (where the enzymes are oriented to the opposite directions from the DNA duplex) showed a smaller yet measureable improvement over the free soluble enzymes. The impact of 4- and 12-bp-long spacers between the DNA-target sites for the 1,2-propanediol and the mevalonate DNA scaffolds was also analyzed [11]. All scaffolds with 4-bp spacers between zinc finger binding sites were less effective than their 12-bp counterparts.

Lee *et al.* [10] constructed scaffold plasmids to position ADB–enzyme fusions every 20 bp (8 bp spacer), 30 bp (18 bp spacer), and 40 bp (28 bp spacer), so that all scaffold-bound enzymes were on the same side of the DNA program in three-dimensional space. The scaffold with the 8-bp spacer sequence was associated with the most efficient L-threonine production, confirming the finding that close proximity of the metabolic enzymes enhances the product synthesis, most likely through substrate channeling.

As demonstrated with the *trans*-resveratrol and the other biosynthetic pathways, the spatial orientation and clustering of the enzymes on the DNA scaffold are important. Due to the predictable nature of the DNA, it is possible to predict the enzyme orientation *in situ* that simplifies designing the DNA scaffold, which is important for larger enzymes that, due to the steric effect, might prevent binding of other enzymes on a DNA scaffold.

12.4.2 Number of DNA Scaffold Repeats

In addition to the length of a spacer, the number of repeats of the DNA scaffold is important for fine-tuning the biosynthetic metabolic pathway.

Conrado *et al.* examined in detail the effect of increasing the number of scaffold repeats. They constructed scaffolds with enzyme–scaffold ratios in range of 40:1 to 1:3 (e.g., $[1:1:1]_1$ to $[1:1:1]_{16}$). A DNA program with DNA-target sequences for each of three-enzyme pathways for producing 1,2-propanediol was placed on the same plasmid as zinc finger chimeras. The best 1,2-propanediol yield was obtained when the number of scaffolds was 4, regardless of the arrangement of the DNA-target sites (see Section 12.4.3), with 12-bp spacers between the binding sites. For DNA scaffolds with 4-bp spacers between the DNA-target sites, the number of repeats played no role [11].

For mevalonate production, which is also a three-enzyme metabolic pathway, the genes encoding the chimeric biosynthetic enzymes were not on the same plasmid with the DNA program, enabling alternations not only throughout the number of scaffold repeats but also with the copy number of plasmids with the DNA scaffold. The largest yield enhancement came from the 16 repeats of the [1:4:2] scaffold. This was followed closely by several of the scaffolds [1:2:2] with 2, 4, or 16 repeats (Figure 12.4). In agreement with the previous results for 1,2-propanediol and mevalonate, a yield enhancement for the *trans*-resveratrol was observed when the number of scaffold repeats was decreased from 16 to 4.

These improvements highlight the ability to impact biosynthesis via simple changes in scaffold design. The number of scaffold repeats can easily be

changed, not only by changing the DNA program repeats on plasmid DNA but also by changing the number of plasmids in a cell. This can be achieved by introducing different origins of replication, from low to high copy number properties. The biosynthesis of a metabolite represents an additional burden for the cell. If the burden is too high for the cell, the production of a metabolite will not lead to the maximal yield. By changing the number of scaffold repeats, we can determine the state where the production yield of the wanted biosynthetic product is maximal.

12.4.3 DNA-Target Site Arrangement

In addition to the length of a spacer between DNA-target sites and the number of DNA scaffold repeats, the stoichiometry of DNA-target sites for individual enzymes forming biosynthetic pathways could also be varied. This is beneficial for biosynthetic pathways with enzymes with different kinetics.

It should be noted that different enzyme arrangements on plasmid DNA are possible. Different architectures are described as, for example, $[E1_a{:}E2_b{:}E3_c]_n$ for a three-enzyme scaffold, whereas a, b, and c describe the enzyme stoichiometry within a single scaffold unit and n is the number of times the scaffold unit is repeated on the plasmid (Figure 12.7a).

For the L-threonine scaffold, Lee *et al.* [10] used the following architectures [1 : 1 : 1], [1 : 1 : 2], [1 : 1 : 3], [1 : 1 : 4] with an increasing number of homodimer TSs

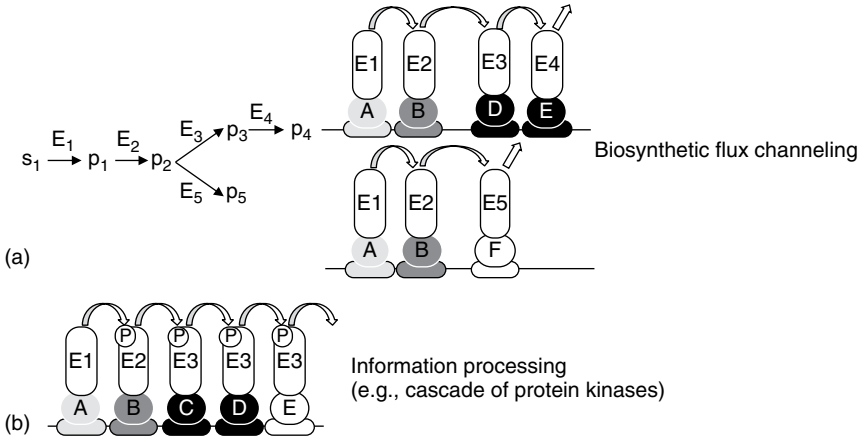


Figure 12.7 Applications of DNA-guided programming. (a) For many biosynthetic pathways, the first enzymes in the cascade are the same, and the end product is determined by the enzymes that are lower in the cascade. By immobilization of specific enzymes on a DNA scaffold, when others are left out, we can determine which end product will be preferentially synthesized. This is a powerful tool for influencing the biosynthetic flux to produce less unwanted products and a cleaner end product. (b) Similar to protein scaffolds, DNA scaffolds can be used for defining the order in which protein kinases phosphorylate each other or a chain of other posttranslational protein modifications. Signaling pathways can, therefore, be modulated using different scaffolds. The DNA scaffold could also be used for information processing such as rewiring intracellular signaling pathways and designing new protein networks for constructing new biological devices with selected features.

in a consecutive manner. The arrangement [1:1:2] with an 8-bp spacer produced the best results, followed closely by [1:1:3], [1:1:4]. They found that the production rate was threefold higher than that of [1:1:1] (Figure 12.2a).

For the 1,2-propanediol and mevalonate synthesis, the scaffolds were designed bidirectionally in the way that the first enzyme was flanked on both sites by the second, followed by the third enzyme [1:2:2]. In addition, a consecutive arrangement of the second enzyme [1:2:1] and [1:4:1] for both 1,2-propanediol and mevalonate biosynthesis, the DkgA and HMGS, respectively, was tested. The DNA scaffold arrangement $[1:2:1]_4$ 12 bp spacer gave the best yield of 1,2-propanediol, closely followed with $[1:2:2]_4$ 12 bp and $[1:4:2]_2$ 12 bp (Figure 12.4). The DNA scaffold [1:4:2] combines both the bidirectional and consecutive arrangement of DNA-target sites and functional units. For mevalonate production, the $[1:4:2]_2$ 12 bp DNA scaffold gave the best yield, followed closely by the $[1:2:2]_{2,4,16}$ 12-bp scaffold [11].

In some biosynthetic pathways, the bottleneck is the conversion rate of a substrate into a product, which can be a substrate for the next enzyme in the metabolic pathway. By changing the arrangement of biosynthetic enzymes on a DNA scaffold, the imbalances in the enzyme kinetics can be overcome. It might be expected that the multimerization of functional enzymes could interfere with the formation of functional scaffolds; however, the biosynthesis of L-threonine depends on enzymes that are active as homotetramers and homodimers, and still, the DNA scaffold improves the production rate of L-threonine [10]. The fact that multimeric proteins might facilitate DNA scaffold cross-linking, therefore building regions with locally elevated concentrations of metabolites (metabolite microdomains), is dedicated for bioconversion. Moon and coworkers [30] showed a positive correlation between a glucaric acid titer and the number of scaffold interaction domains targeting upstream *myo*-inositol-1-phosphate synthase. In the mevalonate pathway, protein scaffolding generating microdomains enabled faster growth rates, likely minimizing the cellular accumulation of the toxic intermediate HMG-CoA in *E. coli* [8].

Taken together, the DNA scaffold is a useful tool to improve biosynthesis. The predictable nature of DNA enables the fine-tuning of metabolic biosynthesis and production yields.

12.5 Applications of DNA-Guided Programming

By studying different DNA scaffold architectures, enzyme stoichiometry, and flux balanced or imbalanced biosynthetic pathways, it should be possible to determine when the enzyme co-localization is most beneficial. This, in turn, will be very useful for guiding the future design of these systems and in envisioning new applications for enzyme co-localization. It is also worth mentioning that the DNA scaffold approach is highly complementary to many of the existing methods for enzyme, pathway, and strain engineering that are already in the cellular engineer's toolkit. Therefore, a successful strategy for achieving the production yields near theoretical maximum is necessary for the commercial viability of production processes and will likely involve a combination of these approaches.

Many biosynthetic pathways are also branched, which means that the enzymes at the initial steps are shared and the enzymes after the branch differs, which determines the end products and their ratios. With DNA scaffolds where the order of enzymes can be changed, the end product could be determined by scaffolding the selected pathway. This leads to the production of cleaner end products and less unwanted products (Figure 12.7a). In addition, with a substrate to product channeling, which is achieved by the DNA scaffold, the accumulation of intermediate products that are toxic for the cell or that can significantly slow down the production rate is consumed faster.

Moreover, the DNA-guided assembly could also be used outside the cell to support biosynthetic reactions *in vitro*, comprising (i) functional units, for example, biosynthetic enzymes linked to DNA-binding domains or linked to single-stranded oligonucleotides by chemical modification [13–15]; (ii) a DNA scaffold comprising one or more target site sequences; and (iii) a substrate for the first enzyme and cofactors for the enzymes provided to the mixture. Erkelenz *et al*. [31] generated a hybrid DNA–protein device based on the two cytochrome P450 BM3 subdomains conjugated to oligonucleotides. The two conjugates arranged on a switchable DNA scaffold form active monooxygenase, which could be turned off by DNA strand displacement.

DNA scaffolds could also be used to control the flow of different classes of biological information mediators that extend beyond the metabolic pathways and small molecule products. For example, DNA scaffolds could be used to rewire intracellular signaling pathways or to coordinate other assembly-line processes, such as protein folding, degradation, and posttranslational modifications (Figure 12.7a,b). Thus, we anticipate that DNA scaffolds should enable the construction of reliable protein networks to program a range of cellular events. Even though the beauty of nature's most elegant compartmentalization strategies, such as a protected tunnel [32] or intracellular organelles [33, 34], has yet to be recapitulated by engineers, the use of DNA scaffolds is an important early step toward this goal.

Strain development is still hampered by the intrinsic inefficiency of enzymatic reactions caused by simple diffusion and the random collision of enzymes and metabolites. Scaffolding strategies that promote the proximity of metabolic enzymes and direct metabolic intermediates through the catalytic assembly steps are promising solutions for the named problem [7–12]. Regardless of scaffold type, the enzyme assembly increases the local concentration of intermediates around the enzyme on the scaffold, preventing the loss of intermediates by competing reactions and overcoming the problem of toxic intermediates due to the rapid conversion of inhibitors.

Definitions

The DNA **scaffold** is a DNA molecule that serves as a platform for the spatial organization of DNA-binding protein domains. The sequential order of the DNA-binding protein domains with their fusion partners is defined through the DNA-target sites positioned along the DNA molecule. The ordering of the

DNA-binding domains consequently defines the order of the enzymes of the metabolic pathway, which are genetically fused to the DNA-binding domains. The overall speed and effectiveness of reaction catalysis can be improved by the presence of a DNA scaffold

A **protein scaffold** has similar characteristics to the DNA scaffold but is protein based. In contrast to the DNA scaffold, where no natural examples are known, protein scaffolds also occur in nature

The DNA-**target site** or DNA-**binding element** is a nucleotide sequence that is recognized by the DNA-binding domain

DNA **program** stands for the defined order of DNA-**target sites** on the DNA **scaffold**. **Spacers** in the DNA sequence separate the DNA-**target sites** on the DNA **program**, which determines the spatial orientation of the enzymes bound to the DNA relative to each other

Substrate channeling is the transfer of a product of one enzyme directly to the next enzyme with minimal release into the bulk solution. The result of substrate channeling is an improved overall reaction efficiency compared to the situations where the enzymes are randomly distributed within the cytoplasm

The synthetic DNA-**binding protein** is a designed protein that binds a predefined DNA sequence. Individual modules of zinc fingers or TAL proteins are used for the construction of synthetic DNA-binding domains. Each module of the zinc finger has a defined specificity for the nucleotide triplet on the DNA molecule. Similarly, each module of the TAL protein can bind a single predefined nucleotide

Spatial organization is a defined arrangement of components in space. Within the context of metabolic engineering, this means that biosynthetic enzymes are fixed in a defined arrangement imposed by the scaffold

The **fusion protein** or **chimeric protein** is a protein created through the joining of two or more genes that code for individual proteins or protein domains. In our case, this refers to the fusion of an enzyme and a DNA-binding domain

References

1 Pfleger, B.F., Pitera, D.J., Smolke, C.D., and Keasling, J.D. (2006) Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. *Nat. Biotechnol.*, **24**, 1027–1032. http://www.ncbi.nlm.nih.gov/pubmed/16845378 (accessed 23 March 2014).

2 Pitera, D.J., Paddon, C.J., Newman, J.D., and Keasling, J.D. (2007) Balancing a heterologous mevalonate pathway for improved isoprenoid production in *Escherichia coli*. *Metab. Eng.*, **9**, 193–207. http://www.ncbi.nlm.nih.gov/pubmed/17239639 (accessed 6 March 2013).

3 Bloom, J.D., Meyer, M.M., Meinhold, P., Otey, C.R., MacMillan, D., and Arnold, F.H. (2005) Evolving strategies for enzyme engineering. *Curr. Opin. Struct. Biol.*, **15**, 447–452. http://www.ncbi.nlm.nih.gov/pubmed/16006119 (accessed 23 March 2014).

4 Steen, E.J., Kang, Y., Bokinsky, G., Hu, Z., Schirmer, A., McClure, A., Del Cardayre, S.B., and Keasling, J.D. (2010) Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature*, **463**, 559–562. http://www.ncbi.nlm.nih.gov/pubmed/20111002 (accessed 23 March 2014).

5 Qian, Z.-G., Xia, X.-X., and Lee, S.Y. (2009) Metabolic engineering of *Escherichia coli* for the production of putrescine: a four carbon diamine. *Biotechnol. Bioeng.*, **104**, 651–662. http://www.ncbi.nlm.nih.gov/pubmed/19714672 (accessed 23 March 2014).

6 Qian, Z.-G., Xia, X.-X., and Lee, S.Y. (2011) Metabolic engineering of *Escherichia coli* for the production of cadaverine: a five carbon diamine. *Biotechnol. Bioeng.*, **108**, 93–103. http://www.ncbi.nlm.nih.gov/pubmed/20812259 (accessed 23 March 2014).

7 Beekwilder, J., Wolswinkel, R., Jonker, H., Hall, R., de Vos, C.H.R., and Bovy, A. (2006) Production of resveratrol in recombinant microorganisms. *Appl. Environ. Microbiol.*, **72**, 5670–5672. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1538726&tool=pmcentrez&rendertype=abstract (accessed 23 March 2014).

8 Dueber, J.E., Wu, G.C., Malmirchegini, G.R., Moon, T.S., Petzold, C.J., Ullal, A.V., Prather, K.L.J., and Keasling, J.D. (2009) Synthetic protein scaffolds provide modular control over metabolic flux. *Nat. Biotechnol.*, **27**, 753–759. http://www.ncbi.nlm.nih.gov/pubmed/19648908 (accessed 23 March 2014).

9 Delebecque, C.J., Lindner, A.B., Silver, P.A., and Aldaye, F.A. (2011) Organization of intracellular reactions with rationally designed RNA assemblies. *Science (New York, N.Y.)*, **333**, 470–474. http://www.ncbi.nlm.nih.gov/pubmed/21700839 (accessed 23 March 2014).

10 Lee, J.H., Jung, S.-C., Bui, L.M., Kang, K.H., Song, J.-J., and Kim, S.C. (2013) Improved production of l-threonine in *Escherichia coli* by use of a DNA scaffold system. *Appl. Environ. Microbiol.*, **79**, 774–782. http://www.ncbi.nlm.nih.gov/pubmed/23160128 (accessed 23 March 2014).

11 Conrado, R.J., Wu, G.C., Boock, J.T., Xu, H., Chen, S.Y., Lebar, T., Turnšek, J., Tomšič, N., Avbelj, M., Gaber, R., Koprivnjak, T., Mori, J., Glavnik, V., Vovk, I., Benčina, M., Hodnik, V., Anderluh, G., Dueber, J.E., Jerala, R., and DeLisa, M.P. (2012) DNA-guided assembly of biosynthetic pathways promotes improved catalytic efficiency. *Nucleic Acids Res.*, **40**, 1879–1889. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3287197&tool=pmcentrez&rendertype=abstract (accessed 21 May 2013).

12 Zhang, Y., Li, S.-Z., Li, J., Pan, X., Cahoon, R.E., Jaworski, J.G., Wang, X., Jez, J.M., Chen, F., and Yu, O. (2006) Using unnatural protein fusions to engineer resveratrol biosynthesis in yeast and Mammalian cells. *J. Am. Chem. Soc.*, **128**, 13030–13031. http://www.ncbi.nlm.nih.gov/pubmed/17017764 (accessed 23 March 2014).

13 Niemeyer, C.M., Koehler, J., and Wuerdemann, C. (2002) DNA-directed assembly of bienzymic complexes from in vivo biotinylated NAD(P)H:FMN oxidoreductase and luciferase. *ChemBioChem*, **3**, 242–245. http://www.ncbi.nlm.nih.gov/pubmed/11921405 (accessed 23 March 2014).

14 Müller, J. and Niemeyer, C.M. (2008) DNA-directed assembly of artificial multienzyme complexes. *Biochem. Biophys. Res. Commun.*, **377**, 62–67. http://www.ncbi.nlm.nih.gov/pubmed/18823945 (accessed 23 March 2014).

15 Wilner, O.I., Weizmann, Y., Gill, R., Lioubashevski, O., Freeman, R., and Willner, I. (2009) Enzyme cascades activated on topologically programmed DNA scaffolds. *Nat. Nanotechnol.*, **4**, 249–254. http://www.ncbi.nlm.nih.gov/pubmed/19350036 (accessed 23 March 2014).
16 Watts, K.T., Lee, P.C., and Schmidt-Dannert, C. (2006) Biosynthesis of plant-specific stilbene polyketides in metabolically engineered *Escherichia coli*. *BMC Biotech.*, **6**, 22. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1435877&tool=pmcentrez&rendertype=abstract (accessed 9 April 2013).
17 Altaras, N.E. and Cameron, D.C. (1999) Metabolic engineering of a 1,2-propanediol pathway in *Escherichia coli*. *Appl. Environ. Microbiol.*, **65**, 1180–1185. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=91161&tool=pmcentrez&rendertype=abstract (accessed 23 March 2014).
18 Martin, V.J.J., Pitera, D.J., Withers, S.T., Newman, J.D., and Keasling, J.D. (2003) Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. *Nat. Biotechnol.*, **21**, 796–802. http://www.ncbi.nlm.nih.gov/pubmed/12778056 (accessed 6 March 2013).
19 Laity, J.H., Lee, B.M., and Wright, P.E. (2001) Zinc finger proteins: new insights into structural and functional diversity. *Curr. Opin. Struct. Biol.*, **11**, 39–46. http://www.ncbi.nlm.nih.gov/pubmed/11179890 (accessed 23 March 2014).
20 Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A., and Bonas, U. (2009) Breaking the code of DNA binding specificity of TAL-type III effectors. *Science (New York, N.Y.)*, **326**, 1509–1512. http://www.ncbi.nlm.nih.gov/pubmed/19933107 (accessed 19 March 2014).
21 Fu, F. and Voytas, D.F. (2013) Zinc Finger Database (ZiFDB) v2.0: a comprehensive database of C_2H_2 zinc fingers and engineered zinc finger arrays. *Nucleic Acids Res.*, **41**, D452–D455. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531203&tool=pmcentrez&rendertype=abstract (accessed 24 March 2014).
22 Greisman, H.A. and Pabo, C.O. (1997) A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science (New York, N.Y.)*, **275**, 657–661. http://www.ncbi.nlm.nih.gov/pubmed/9005850 (accessed 23 March 2014).
23 Rebar, E.J. and Pabo, C.O. (1994) Zinc finger phage: affinity selection of fingers with new DNA-binding specificities. *Science (New York, N.Y.)*, **263**, 671–673. http://www.ncbi.nlm.nih.gov/pubmed/8303274 (accessed 23 March 2014).
24 Maeder, M.L., Thibodeau-Beganny, S., Osiak, A., Wright, D.A., Anthony, R.M., Eichtinger, M., Jiang, T., Foley, J.E., Winfrey, R.J., Townsend, J.A., Unger-Wallace, E., Sander, J.D., Müller-Lerch, F., Fu, F., Pearlberg, J., Göbel, C., Dassie, J.P., Pruett-Miller, S.M., Porteus, M.H., Sgroi, D.C., Iafrate, A.J., Dobbs, D., McCray, P.B., Cathomen, T., Voytas, D.F., and Joung, J.K. (2008) Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Mol. Cell*, **31**, 294–301. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2535758&tool=pmcentrez&rendertype=abstract (accessed 23 March 2014).
25 Hurt, J.A., Thibodeau, S.A., Hirsh, A.S., Pabo, C.O., and Joung, J.K. (2003) Highly specific zinc finger proteins obtained by directed domain shuffling and

cell-based selection. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 12271–12276. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=218748&tool=pmcentrez&rendertype=abstract (accessed 23 March 2014).

26 Pavletich, N.P. and Pabo, C.O. (1991) Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science (New York, N.Y.)*, **252**, 809–817. http://www.ncbi.nlm.nih.gov/pubmed/2028256 (accessed 23 March 2014).

27 Yang, W.-P.P., Wu, H., and Barbas, C.F.F. (1995) Surface plasmon resonance based kinetic studies of zinc finger-DNA interactions. *J. Immunol, Methods*, **183**, 175–182. http://www.ncbi.nlm.nih.gov/pubmed/7602135 (accessed 23 March 2014).

28 Stains, C.I., Porter, J.R., Ooi, A.T., Segal, D.J., and Ghosh, I. (2005) DNA sequence-enabled reassembly of the green fluorescent protein. *J. Am. Chem. Soc.*, **127**, 10782–10783. http://www.ncbi.nlm.nih.gov/pubmed/16076155 (accessed 23 March 2014).

29 Garg, A., Lohmueller, J.J., Silver, P.A., and Armel, T.Z. (2012) Engineering synthetic TAL effectors with orthogonal target sites. *Nucleic Acids Res.*, **40**, 7584–7595. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3424557&tool=pmcentrez&rendertype=abstract (accessed 23 March 2014).

30 Moon, T.S., Dueber, J.E., Shiue, E., and Prather, K.L.J. (2010) Use of modular, synthetic scaffolds for improved production of glucaric acid in engineered *E. coli*. *Metab. Eng.*, **12**, 298–305. http://www.ncbi.nlm.nih.gov/pubmed/20117231 (accessed 23 March 2014).

31 Erkelenz, M., Kuo, C.-H., and Niemeyer, C.M. (2011) DNA-mediated assembly of cytochrome P450 BM3 subdomains. *J. Am. Chem. Soc.*, **133**, 16111–16118. 10.1021/ja204993s (accessed 23 March 2014).

32 Hyde, C.C., Ahmed, S.A., Padlan, E.A., Miles, E.W., and Davies, D.R. (1988) Three-dimensional structure of the tryptophan synthase alpha 2 beta 2 multienzyme complex from *Salmonella typhimurium*. *J. Biol. Chem.*, **263**, 17857–17871. http://www.ncbi.nlm.nih.gov/pubmed/3053720 (accessed 23 March 2014).

33 Bobik, T.A. (2006) Polyhedral organelles compartmenting bacterial metabolic processes. *Appl. Microbiol. Biotechnol.*, **70**, 517–525. http://www.ncbi.nlm.nih.gov/pubmed/16525780 (accessed 23 March 2014).

34 Straight, P.D., Fischbach, M.A., Walsh, C.T., Rudner, D.Z., and Kolter, R. (2007) A singular enzymatic megacomplex from *Bacillus subtilis*. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 305–310. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1765455&tool=pmcentrez&rendertype=abstract (accessed 22 April 2013).

13

Synthetic RNA Scaffolds for Spatial Engineering in Cells

Gairik Sachdeva[*,1,2,3], *Cameron Myhrvold*[*,2,3], *Peng Yin*[2], *and Pamela A. Silver*[2,3]

[1] *Harvard John A. Paulson School of Engineering and Applied Sciences, 29 Oxford Street, Cambridge, MA 02138, USA*
[2] *Harvard University, Wyss Institute for Biologically Inspired Engineering, 3 Blackfan Circle, Boston, MA 02115, USA*
[3] *Harvard Medical School, Department of Systems Biology, 200 Longwood Avenue, Boston, MA 02115, USA*

13.1 Introduction

The ability to engineer cells with subcellular spatial precision is a very powerful and essential tool in synthetic biology. Specifically, co-localization of proteins, DNA, and RNA enhances metabolic output of enzymes [1, 2], allows novel regulation of gene expression [3–5], and can increase the specificity of therapeutics [6, 7]. This occurs primarily because co-localized macromolecules have high local concentrations, allowing their activities to be coordinated. Thus, better ability to organize proteins, RNAs, lipids, etc. into synthetic macromolecular complexes should enable diverse and more complex function than can be achieved by solely engineering individual parts.

In this chapter, we illustrate how synthetic RNA constructs are advancing efforts toward *in vivo* spatial engineering. Natural noncoding RNAs already play structural and catalytic roles in cells. A breadth of studies has established design principles that can be used to predictably shape RNA secondary structures [8–11]. Structural malleability of RNA, the ease of expressing synthetic RNA constructs in cells, their stability, and advances in methods for assaying and imaging assembled structures are some of the many reasons why RNA is a useful scaffolding material. Synthetic biology efforts have demonstrated that carefully designed RNA can be used for subcellular targeting of probes, enzymes, and therapeutic agents.

13.2 Structural Roles of Natural RNA

RNAs perform numerous biological functions as canonical gene expression agents, catalysts, gene regulation switches, and structural scaffolds. These struc-

* These authors contributed equally to the work

Synthetic Biology: Parts, Devices and Applications, First Edition. Edited by Christina Smolke.

tural and catalytic roles of RNA are due in large part to the tremendous diversity of secondary and tertiary structures assumed by natural RNA and the fact that ribose sugars are more reactive than deoxyribose. RNA secondary structures can include intricate motifs like double helices, hairpin loops, bulges, pseudoknots, and right-angled turns [12, 13]. Aside from the Watson–Crick base pairing, RNA has the capacity to form Hoogsteen base pairs as well as wobble base pairs. Such interactions allow motifs to be connected in higher-order tertiary interactions, predominantly through the non-Watson–Crick base pairs [14, 15].

13.2.1 RNA as a Natural Catalyst

Catalytic roles of RNA during translation, like the tRNA shown in Figure 13.1a, disrupted a simple view held by the central dogma that RNA exists merely to transfer genetic information from DNA to protein. Today we know that RNA has catalytic and regulatory roles in many other cellular processes as well. Regulatory RNA structures play a significant role in the control of translation initiation of several bacterial genes and in bacterial immunity [17]. RNAs affect expression in *cis*, by forming secondary structures near translation start sites of the mRNA. The *cis* regions can bind to regulatory proteins or other RNAs that affect translation in *trans* [17]. Other similarly dynamic regulatory RNA regions can consist of aptamers, which are nucleic acids that selectively bind ligands

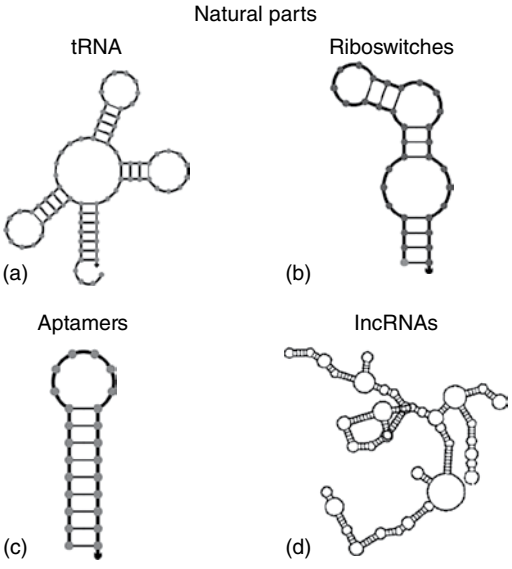


Figure 13.1 Prevalence and diversity of secondary structure in natural RNA. (a) The alanine-carrying transfer RNA shown here has the typical clover leaf structure common among tRNA. (b) The theophylline-binding riboswitch (from PDB: 1O15_A) is a canonical riboswitch. (c) The PP7 aptamer [16] binds to the PP7 coat protein with low nanomolar affinity. (d) The *Homo sapiens* TERC lncRNA (NR_001566.1) is an example of a natural lncRNA that serves as a scaffold.

[18]. Many metabolic genes are "switched" on or off, triggered by the binding of small molecule metabolites to some of these regulatory RNAs known as riboswitches (Figure 13.1b) [19].

13.2.2 RNA Scaffolds in Nature

There are also several instances of natural RNAs that are largely structural. Some natural RNAs are known to specifically bind the coat proteins of single-strand RNA phages. Such interactions help package the RNA into viral capsids. Some RNA phages that have well-characterized RNA-binding proteins include PP7 (Figure 13.1c) [16], MS2 [20], and Qβ [21]. These coat proteins also act as repressors of the viral replicase translation by specifically binding RNA hairpins near the origin of replication. In the bacteriophage Φ29, a short (117–174 nt) sequence of packaging RNA (pRNA) helps to pack phage DNA into preformed capsids [22]. A DNA packaging motor is composed of a pentameric ring of pRNA, capsid proteins, dsDNA, and an ATPase [23]. Studies characterizing the specificity and stoichiometries of these interactions [16, 24–26] have laid the foundation for RNA-tagging-based applications that we look at in Section 13.4.

RNA scaffolds are important in eukaryotic gene expression as well. Mammalian cells appear to extensively employ long noncoding RNAs (lncRNAs). These lncRNAs (Figure 13.1d) are rich with secondary structure motifs [27, 28], some of which bind and coordinate proteins on scaffolds that play important roles in epigenetic regulation [29, 30] and telomere maintenance [31, 32].

Thus, natural RNA diversity offers a template of diverse structure and function for synthetic biologists. In the following section, we look at how natural observations have been translated into an understanding of the means to precisely engineer structure and dynamics of RNA.

13.3 Design Principles for RNA Are Well Understood

In order to design, build, and test structures at the molecular scale, one must understand the physical properties of the building material. In particular, if one uses a biopolymer such as a protein or nucleic acid to build a higher-order structure, the folding properties of that polymer will dictate the structure. This is especially a challenge in the case of protein engineering, where protein structure is extremely difficult to predict *ab initio* [33, 34]. As a result, many protein engineers have focused on substituting functional rather than structural residues in existing proteins [35]. Unlike proteins, nucleic acids have a well-defined helical structure governed by a simple set of complementarity rules [36] with some exceptions such as wobble pairing and G quadruplexes [37, 38]. As a result, the structural and folding properties of RNA are generally well understood. In addition, RNA is a dynamic molecule [39–42] that can self-assemble into structures *in vitro* [13, 43–46] and can be easily transcribed from a DNA template *in vivo*. RNA functionality can also be improved using *in vitro* selection [47, 48]. For these reasons, RNA makes a suitable material for constructing synthetic *in vivo* nanostructures.

13.3.1 RNA Secondary Structure is Predictable

Most RNAs fold into a secondary structure consisting of a series of base-paired stems and unpaired loops. This secondary structure is largely determined by complementary bases within the primary RNA sequence. As a result, RNA secondary structure can be predicted computationally using a variety of methods. This typically involves using a model of the free energy of RNA base pairing [49, 50] to determine the minimum free energy secondary structure [8–11]. Structures with near-optimal folds are also calculated by these software packages, since they may be of interest, and partition functions are used to determine the relative probabilities of particular secondary structures based on their energetics (Figure 13.2a) [8–11]. Additional factors, such as wobble base pairing, pseudoknots, and dangling bases, are often incorporated into these calculations [8, 55].

Several software packages have been developed for the purpose of calculating DNA or RNA secondary structure. These include UNAFold, RNAstructure, NUPACK, and ViennaRNA [8–10, 55]. The software is typically implemented as a web server that can be used to run calculations using an online interface; it is also possible to install a local copy of the software. Each package has a somewhat different feature set (see Table 13.1 for details). For example, RNAstructure can

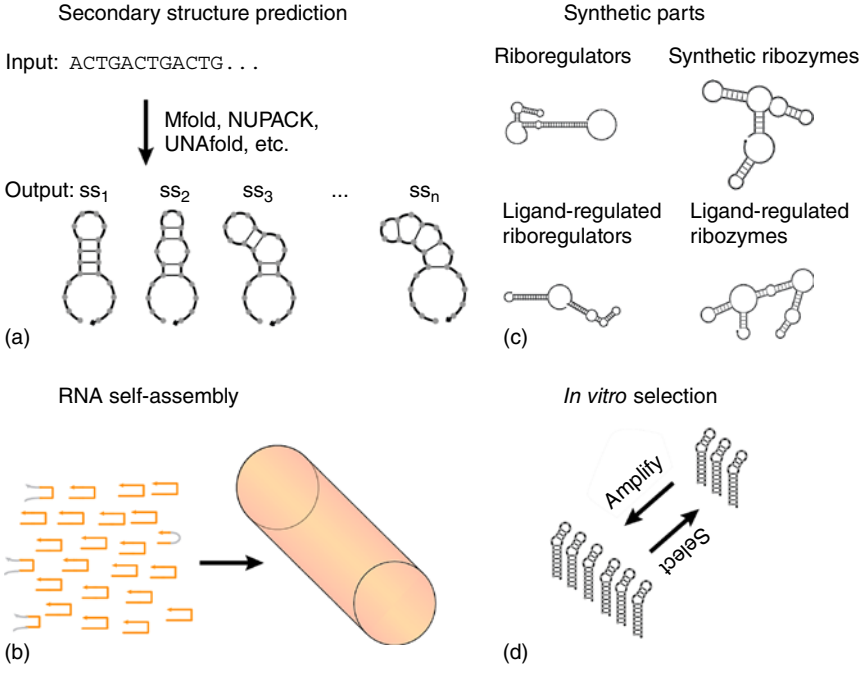


Figure 13.2 Design principles for RNA structure and function. (a) RNA secondary structure can be predicted from the primary sequence using a variety of software packages. (b) RNA can self-assemble into 2D or 3D structures *in vitro*. (c) Researchers have developed a variety of synthetic parts, such as synthetic riboregulators, synthetic ribozymes, ligand-regulated riboregulators, and ligand-regulated ribozymes [51–54]. (d) *In vitro* selection can be used to enhance the function of RNAs through iterative rounds of amplification and selection.

Table 13.1 Comparison of features between RNA structure prediction software packages.

Feature	NUPACK	RNAstructure	UNAfold	ViennaRNA
MFE calculation	●	●	●	●
Partition function	●	●	●	●
Wobble pairing	●	●	●	●
Pseudoknots	●	●	○	○
Dangling bases	●	●	●	●
Multi-strand interactions	●	○	○	○
Uses SHAPE/NMR data	○	●	○	○
Graphical User Interface	○	●	○	○
Web Interface	●	●	●	●

A filled-in circle indicates that the software package contains the feature in a row, whereas an empty circle indicates that the software package does not contain the feature in a row.
MFE, minimum free energy.

integrate user-supplied experimental data such as selective 2′ hydroxylation and primer extension (SHAPE) [56] or NMR to aid in structure calculation and has a convenient graphical user interface [10]. ViennaRNA is designed to be computationally efficient for testing many RNA structures in batches rather than for analyzing individual species in more detail [8]. UNAFold is derived from mfold, which used the first dynamic programming algorithm for predicting RNA secondary structure [9, 57]. A particularly useful package for designing RNA structures is NUPACK, which can handle multi-strand interactions and allows the user to design sequences that have a propensity to assemble into a user-defined set of secondary structures [55, 58]. Given the diversity of software packages for predicting RNA secondary structure, it is important to choose the right software package for one's particular design needs.

13.3.2 RNA can Self-Assemble into Structures

RNA can self-assemble into geometrically precise structures *in vitro* (Figure 13.2b). This was first shown for small RNA molecules with four stem-loops (tectoRNAs), which self-assemble into 1D structures using kissing loops [59], but has since been extended to form a variety of geometrically precise 2D and 3D shapes [13, 43–46, 60]. Of particular note are the *in vivo* RNA assemblies [1], which can self-assemble into 1D or 2D lattices. Although *in vitro* structures have traditionally been formed using a thermal annealing process, recent work has shown that single-stranded DNA tiles and bricks [61, 62] can self-assemble into discrete nanostructures isothermally and under biocompatible conditions [63]. Thus, it is possible to self-assemble a diverse range of scaffolds using RNA.

13.3.3 Dynamic RNAs can be Rationally Designed

Beyond structure formation, RNA also has the capability to dynamically reconfigure itself in response to small molecules or other ligands [39–42]. Such

RNAs – ribozymes and riboswitches, respectively – underscore the notion that RNAs can be dynamic molecules. However, RNAs can also be rationally designed to go beyond their natural function (Figure 13.2c). For example, synthetic riboregulators can be designed to control genes in the presence of a user-defined input RNA molecule [51]. It is even possible to combine pairs of functional RNAs to form more complicated devices, such as by combining riboswitches with ribozymes [64], riboswitches with riboregulators [52, 65], or aptamers with transcriptional attenuators [66]. These compound RNA devices underscore the notion that RNA secondary structure can be programmed to achieve a range of dynamic functions.

13.3.4 RNA can be Selected *in vitro* to Enhance Its Function

Another powerful technique that has aided the development of many functional RNA motifs is *in vitro* selection or systematic evolution of ligands by exponential enrichment (SELEX) [47, 48] (Figure 13.2d). This typically involves starting with a library of many (10^{13}–10^{15}) distinct RNA sequences and then applying iterative rounds of selection (e.g., binding to a small molecule immobilized on a surface or catalyzing ligation to a surface-bound ligand) and amplification (typically involving polymerase chain reaction (PCR)). After ~10 rounds of selection and amplification, the activity of the remaining RNA sequences in the pool can be enhanced by several orders of magnitude compared with the initial library average [67]. Some functions may not be present in a library of 10^{15} RNAs; thus it may sometimes be necessary to chemically modify or structurally bias the initial library [67]. This limitation aside, *in vitro* selection is a useful technique for generating synthetic RNAs with specific functions.

In the two decades since the development of *in vitro* selection, thousands of aptamers (oligonucleotides that bind to a particular ligand) have been developed [68]. These include aptamers to small molecules, peptides, and even human and cancer cell types [47, 67, 69–71]. In addition to RNA molecules, proteins such as epitopes and antibodies have been evolved using *in vitro* selection [72–74]. Thus, *in vitro* selection can be used to enhance functional portions of an RNA scaffold. This is especially useful when existing RNA parts are not sufficient for the task at hand.

13.4 Applications of Designed RNA Scaffolds

RNA sequences consisting of secondary structures and functional units designed using the tools described previously can be genetically expressed in cells. Such engineered RNAs have been used for tasks ranging from studying natural RNA processing in cells to metabolic engineering and therapeutic applications.

13.4.1 Tools for RNA Research

While mRNA has long been known to function as a template for protein translation, the spatiotemporal aspects of the various steps involved in mRNA processing

remain poorly understood. Investigation of the dynamics of mRNA as it goes through translation, splicing, nuclear export in eukaryotes, localization for translation, and finally degradation requires tools to track individual RNA molecules. Aptamers and their recruitment of fluorescent proteins on engineered mRNA scaffolds have enabled such studies.

Some of the earliest attempts to tag RNA *in vivo* were carried out by expressing GFP fused with bacteriophage MS2 coat protein [75] or human RNA-interacting protein domain U1A [76] along with mRNA containing the corresponding binding sites in *Saccharomyces cerevisiae*. Such tags enabled tracking of single-cell mRNA localization by microscopy. Furthermore, by incorporating tandem repeats of MS2 binding sites on reporter mRNA [77], several GFP–MS2 fusions could be localized on a single transcript, enabling tracking of individual mRNA molecules in mammalian cells (Figure 13.3a). This *in vivo* tracking method was extended to other systems [82], including bacteria [78, 83].

More recently, several efforts have addressed the long-standing question of whether or not RNA is highly localized within bacterial cells [84, 85]. A significant innovation over the previous strategy came from the use of fluorescent protein complementation. In this approach, RNA aptamers are used to bring together two different protein fusion units, each with a split fluorescent protein fused to an RNA-binding domain (RBD) [79, 86] (Figure 13.3a). Since only the scaffolded protein units are able to reconstitute the split chromophore and fluoresce, they can be easily distinguished from the unbound ones. Such an approach hence achieves lower background signals than systems where autofluorescent proteins are directly tagged onto RNA.

As the repertoire of aptamer–RNA-binding protein pairs is being extended through the *in vitro* methods described in Section 13.3.4, newer combinations are being used to explore cellular function [87]. The studies discussed here have led to a better understanding of RNA diffusion and localization [78, 79] in bacterial cells and measurement of transcriptional kinetics [88]. These efforts also enabled localization of a diverse array of proteins (such as enzymes) on RNA scaffolds, opening up applications in metabolic engineering.

13.4.2 Localizing Metabolic Enzymes on RNA

Scaffolding and compartmentalization are effective strategies for optimization of metabolic pathway performance in both natural and synthetic systems [89, 90]. A few studies have used DNA structures to coordinate the assembly of enzymes and study effects of spatial co-localization *in vitro* [91–94] and *in vivo* [95]. Protein scaffolds have also been used to channel metabolic substrates between co-localized enzymes in living cells [2, 96]. Scaffolding is seen as a powerful tool to specifically direct metabolic pathway flux toward enzymes of choice, prevent loss of intermediates to competing reactions, and protect the host cell from any toxic or volatile intermediates through confinement at a subcellular location.

A notable effort in the use of RNA scaffolds for metabolic channeling achieved a nearly 50-fold increase in hydrogen gas production in *Escherichia coli* [1]. This effort combined many of the techniques discussed previously. Synthetic RNA

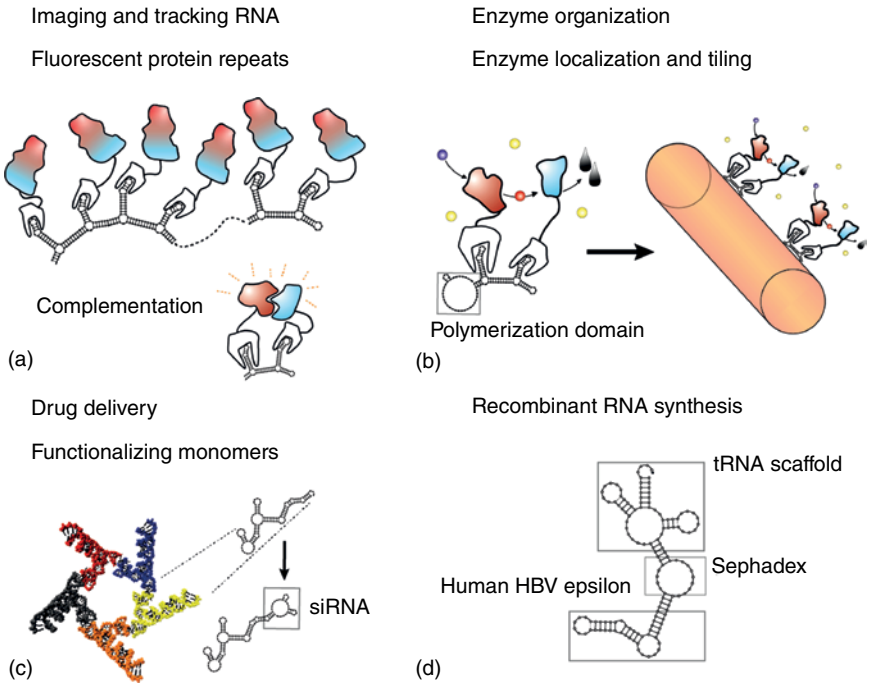


Figure 13.3 Applications of RNA scaffolds *in vivo*. (a) mRNA are modified to include either several repeats of an aptamer or two different aptamers in close proximity. The former approach results in concentrated foci of fluorescent protein fusions to RNA-binding domains (RBDs) [78] and in the latter, two halves of the protein with RBD fusions [79], only complement to be fluorescent on the mRNA scaffold. (b) Enzymes fused to RBDs localize to self-assembled RNA scaffolds with aptamers presented. Channeling of intermediate metabolites can lead to enhanced pathway flux toward biofuels or other high value products [1]. (c) Pentamer of bacteriophage Φ29 pRNA [23] from PDB file 1FOQ. Tagging the monomers with functional units like siRNA can make them useful drug delivery vehicles [6, 80]. (d) The clover leaf tRNA sequence can be tagged with recombinant RNA and epitopes as shown to allow for its synthesis and purification [81].

strands comprising polymerization domains and aptamers for MS2 and PP7 coat proteins were expressed in the bacteria. Dimerization and polymerization domains allowed for tiling and assembly into a macromolecular structure. The large (40–100 nm) intracellular RNA assemblies greatly enhanced the flux of electrons from ferredoxin to hydrogenase when both enzymes were tethered to the scaffold with fusions to MS2 and PP7 (Figure 13.3b). Furthermore, significant differences in titer were observed for scaffolding structures having different geometries, tying metabolic flux to the specific spatial positioning of the scaffold. Such an approach brings modular design and scalability [97] to metabolic engineering for biofuels and high value chemical synthesis, where control of intermediate metabolite flux can be critical [98–100].

There has been debate about the mechanism by which scaffolds enable metabolic substrate channeling. The transfer of electrons between enzymes relies on physical contact and thus is limited by protein diffusion rates and competition,

which are effectively addressed by scaffolding [1]. However, the role of enzyme co-localization in pathways involving diffusible intermediates is much less well understood [101, 102]. In a recent study [103], enzymes localized in close proximity, less than 30 nm apart, on *in vitro* assembled DNA scaffolds exhibited enhanced rates of metabolite exchange. The transfer rates dropped precipitously with any further increase in interenzyme distance. Since such effects are not explicable by 3D diffusion models [101], a mechanism of metabolite substrate channeling by restricted diffusion on hydration layers across crowded protein surfaces has been proposed [103]. RNA scaffolds, with their predictable geometry, can be used to create a range of metabolic channeling platforms and test the relative effects from these two different mechanisms.

13.4.3 Packaging Therapeutics on RNA Scaffolds

While metabolic channeling functions relied on RNA interactions with proteins, RNA–RNA interactions can also be used for exciting scaffold applications. pRNA from bacteriophage Φ29 (referred to in Section 13.2) has been used as a building block for bottom-up assembly of drug delivery vehicles [6, 80] (Figure 13.3c). pRNA monomers consist of structural hairpin regions and dimerization/polymerization domains. Ends of the hairpin regions offer sites for tagging with drugs or targeting molecules. The polymerization domains can be engineered to favor formation of dimers, trimers, pentamers, or hexamers as stable drug carriers [6, 23, 80]. Heterodimers containing pRNA tagged with a CD4 aptamer and pRNA attached to an siRNA were shown to specifically target CD4-expressing T cells, leading to cell death [80]. This *in vitro* study also showed stability and efficacy of the nanoscale drug delivery particles for killing cancer cells. Such systems are advantageous since the pRNA polymers are hypothesized to be stable in physiological conditions and be less immunogenic than protein carriers [80]. Finally, these polymers could be made specific to many *in situ* targets by using engineered specific RNA aptamers that recognize cellular moieties.

13.4.4 Recombinant RNA Technology

RNA scaffolds have also been used to serve as protective tethers for the purification of recombinant RNA (recRNA) (Figure 13.3d) [81]. In this approach, a tRNA scaffold acts as a protective secondary structure to insulate the transcript from native *E. coli* nucleases and therefore stabilize production of recRNA *in vivo*. The characteristic clover leaf tRNA structure formed around a recRNA is recognized by native cellular enzymes and processed as tRNA. This ensures that each single transcript is a product of specific defined length. A Sephadex affinity tag was included in the expressed sequence to allow purification of transcripts that contained RNAs of medical research interest, like the human hepatitis B virus (HBV) epsilon [81]. This design thus enables collection of large amounts of purified RNA transcripts for *in vitro* structural studies and vaccine development. Recently, these efforts have been extended to expression and purification of RNA–protein complexes [104], providing pure samples that could be used for crystallographic studies of natural RNA–protein interactions and potential use in cell-free systems.

13.5 Conclusion

RNA is a powerful tool to synthetic biologists. RNA scaffolds can be composed of many structural, dynamic, and functional regions. Structure design can be predicted reliably, and there are a growing number of assays for proper structure assembly. In addition, recent advances in DNA construction [105, 106] have made it faster and easier to test new structure designs *in vivo*. Prediction and design of RNA structure in three dimensions remains a challenge. The difficulty of going from a secondary structure design to precise orientation of tertiary scaffold units needs to be addressed for metabolic engineering and therapeutic applications. Additionally, although localization of fluorophores to RNA enables *in vivo* imaging, resolution limits have prevented elucidation of precise geometric details in RNA scaffolds and assemblies within cells. Future technical advances could enable many scientists to construct new RNA scaffolds for a wide range of purposes. In the following text, we discuss a particular set of exciting applications and the technologies that will enable them.

13.5.1 New Applications

Synthetic biologists are constantly seeking to increase the complexity of their devices. RNA synthetic biology is offering tools to enable such control [107]. One particular goal is the construction of orthogonal ribosomes [108], capable of incorporating nonnatural amino acids wherein altered tRNA–protein interactions enable an expanded genetic code [109]. RNA scaffolds are also being employed to devise more precise genome editing tools [110]. For metabolic engineering applications, RNA scaffolds are enabling control over the relative geometric orientations of enzymes in a co-localized pathway, which can lead to better channeling of volatile intermediate metabolites [111]. Therapeutic applications of *in vivo* RNA scaffolds include functionalizing natural RNA scaffolds to enable drug delivery or isolation of pure samples. Similar developments in the fields of DNA packaging and origami for drug delivery [112, 113] could offer strong synergistic opportunities for clinically applicable technologies to be implemented. More generally, the ability to simulate and predict the dynamics of structure-receptor binding interactions should enhance the design of such therapeutics [114].

13.5.2 Technological Advances

Moving forward, innovations in high-throughput design, synthesis, and assaying functions for RNA structures will enable a greater range of applications to be developed. *In silico* design software packages are continuously improving their capabilities, making it possible to computationally generate increasingly complicated structures [55]. In addition to the advances for *in vivo* synthesis and purification of RNAs mentioned previously, developments in chip-based synthesis could enable hundreds of RNA designs to be synthesized *in vitro* at a time [106, 115]. This, coupled with new structure assembly assays such as SHAPE-Seq [116] and improved genetically encodable electron microscopy tags [117, 118], will greatly simplify the testing of more complicated structures. Developments in

RNA imaging [119] can be further advanced by incorporation of docking sites that allow RNA to be probed with oligonucleotides using methods like DNA-PAINT [120], leading to super-resolution imaging *in situ*.

Thus, the discovery of a variety of natural RNA structures and functions, an ever-increasing understanding of how such features can be designed, and an ability to rapidly implement and test ideas are indicators of a significant role for RNA scaffolds in future synthetic biology applications.

Definitions

Synthetic biology is a discipline that seeks to control biology using the principles of engineering

Nanotechnology is the manipulation of matter at the atomic, molecular, and supramolecular scale

RNA scaffolds are macromolecular structures or assemblies of RNA with well-defined secondary structure motifs for spatially organizing other biomolecules. These are typically expressed in living cells for metabolic engineering purposes

Isothermal assembly is a self-assembly of structures at a constant temperature

Metabolic engineering is the production of small molecules or short peptides through the engineering of metabolic pathways

Aptamers are nucleic acid oligonucleotides that bind a specific small molecule or other ligand

References

1 Delebecque, C.J., Lindner, A.B., Silver, P.A., and Aldaye, F.A. (2011) Organization of intracellular reactions with rationally designed RNA assemblies. *Science*, **333** (6041), 470–474.

2 Dueber, J.E., Wu, G.C., Malmirchegini, G.R., Moon, T.S., Petzold, C.J., Ullal, A.V., Prather, K.L.J., and Keasling, J.D. (2009) Synthetic protein scaffolds provide modular control over metabolic flux. *Nat. Biotechnol.*, **27** (8), 753–759.

3 Isaacs, F.J., Dwyer, D.J., and Collins, J.J. (2006) RNA synthetic biology. *Nat. Biotechnol.*, **24** (5), 545–554.

4 Culler, S.J., Hoff, K.G., and Smolke, C.D. (2010) Reprogramming cellular behavior with RNA controllers responsive to endogenous proteins. *Science*, **330** (6008), 1251–1255.

5 Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., and Lim, W.A. (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, **152** (5), 1173–1183.

6 Khaled, A., Guo, S., Li, F., and Guo, P. (2005) Controllable self-assembly of nanoparticles for specific delivery of multiple therapeutic molecules to cancer cells using RNA nanotechnology. *Nano Lett.*, **5** (9), 1797–1808.

7 Aldaye, F.A., Senapedis, W.T., Silver, P.A., and Way, J.C. (2010) A structurally tunable DNA-based extracellular matrix. *J. Am. Chem. Soc.*, **132** (42), 14727–14729.

8 Lorenz, R., Bernhart, S.H., Siederdissen, C.H.Z., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6** (1), 26.

9 Markham, N.R. and Zuker, M. (2008) UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol.*, **453**, 3–31.

10 Reuter, J.S. and Mathews, D.H. (2010) RNA structure: software for RNA secondary structure prediction and analysis. *BMC Bioinf.*, **11** (1), 129.

11 Zadeh, J.N., Steenberg, C.D., Bois, J.S., Wolfe, B.R., Pierce, M.B., Khan, A.R., Dirks, R.M., and Pierce, N.A. (2010) NUPACK: analysis and design of nucleic acid systems. *J. Comput. Chem.*, **32** (1), 170–173.

12 Leontis, N.B., Lescoute, A., and Westhof, E. (2006) The building blocks and motifs of RNA architecture. *Curr. Opin. Struct. Biol.*, **16** (3), 279–287.

13 Jaeger, L. and Chworos, A. (2006) The architectonics of programmable RNA and DNA nanostructures. *Curr. Opin. Struct. Biol.*, **16** (4), 531–543.

14 Cruz, J.A. and Westhof, E. (2009) The dynamic landscapes of RNA architecture. *Cell*, **136** (4), 604–609.

15 Tinoco, I. Jr. and Bustamante, C. (1999) How RNA folds. *J. Mol. Biol.*, **293** (2), 271–281.

16 Lim, F.F., Downey, T.P.T., and Peabody, D.S.D. (2001) Translational repression and specific RNA binding by the coat protein of the Pseudomonas phage PP7. *J. Biol. Chem.*, **276** (25), 22507–22513.

17 Waters, L.S. and Storz, G. (2009) Regulatory RNAs in bacteria. *Cell*, **136** (4), 615–628.

18 Winkler, W.C. and Breaker, R.R. (2005) Regulation of bacterial gene expression by riboswitches. *Annu. Rev. Microbiol.*, **59**, 487–517.

19 Nudler, E. and Mironov, A.S. (2004) The riboswitch control of bacterial metabolism. *Trends Biochem. Sci*, **29** (1), 11–17.

20 Hirao, I., Spingola, M., Peabody, D., and Ellington, A.D. (1998) The limits of specificity: an experimental analysis with RNA aptamers to MS2 coat protein variants. *Mol. Diversity*, **4** (2), 75–89.

21 Witherell, G.W. and Uhlenbeck, O.C. (1989) Specific RNA binding by Q.beta. coat protein. *Biochemistry*, **28** (1), 71–76.

22 Guo, P., Erickson, S., and Anderson, D. (1987) A small viral RNA is required for in vitro packaging of bacteriophage phi 29 DNA. *Science*, **236** (4802), 690–694.

23 Simpson, A.A., Tao, Y., Leiman, P.G., Badasso, M.O., He, Y., Jardine, P.J., Olson, N.H., Morais, M.C., Grimes, S., Anderson, D.L., Baker, T.S., and Rossmann, M.G. (2000) Structure of the bacteriophage phi29 DNA packaging motor. *Nature*, **408** (6813), 745–750.

24 Ni, C.-Z., Syed, R., Kodandapani, R., Wickersham, J., Peabody, D.S., and Ely, K.R. (1995) Crystal structure of the MS2 coat protein dimer: implications for RNA binding and virus assembly. *Structure*, **3** (3), 255–263.

25 Peabody, D.S. and Ely, K.R. (1992) Control of translational repression by protein–protein interactions. *Nucleic Acids Res.*, **20** (7), 1649–1655.

26 Guo, P., Zhang, C., Chen, C., Garver, K., and Trottier, M. (1998) Inter-RNA interaction of phage ϕ29 pRNA to form a hexameric complex for viral DNA transportation. *Mol. Cell*, **2** (1), 149–155.

27 Underwood, J.G., Uzilov, A.V., Katzman, S., Onodera, C.S., Mainzer, J.E., Mathews, D.H., Lowe, T.M., Salama, S.R., and Haussler, D. (2010) FragSeq:

transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods*, **7** (12), 995–1001.

28 Kertesz, M., Wan, Y., Mazor, E., Rinn, J.L., Nutter, R.C., Chang, H.Y., and Segal, E. (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467** (7311), 103–107.

29 Mercer, T.R. and Mattick, J.S. (2013) Structure and function of long noncoding RNAs in epigenetic regulation. *Nat. Struct. Mol. Biol.*, **20** (3), 300–307.

30 Tsai, M.C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E., and Chang, H.Y. (2010) Long noncoding RNA as modular scaffold of histone modification complexes. *Science*, **329** (5992), 689–693.

31 Zappulla, D.C. and Cech, T.R. (2004) Yeast telomerase RNA: a flexible scaffold for protein subunits. *Proc. Natl. Acad. Sci. U.S.A.*, **101** (27), 10024–10029.

32 Theimer, C.A. and Feigon, J. (2006) Structure and function of telomerase RNA. *Curr. Opin. Struct. Biol.*, **16** (3), 307–318.

33 Arnold, F.H. (2001) Combinatorial and computational challenges for biocatalyst design. *Nature*, **409** (6817), 253–257.

34 Bonneau, R. and Baker, D. (2001) Ab initio protein structure prediction: progress and prospects. *Annu. Rev. Biophys. Biomol. Struct.*, **30**, 173–189.

35 Bornscheuer, U.T., Huisman, G.W., Kazlauskas, R.J., Lutz, S., Moore, J.C., and Robins, K. (2012) Engineering the third wave of biocatalysis. *Nature*, **485** (7397), 185–194.

36 Watson, J.D. and Crick, F.H. (1953) Molecular structure of nucleic acids. *Nature*, **171** (4356), 737–738.

37 Varani, G. and McClain, W.H. (2000) The G· U wobble base pair. *EMBO Rep.*, **1** (1), 18–23.

38 Lipps, H.J. and Rhodes, D. (2009) G-quadruplex structures: in vivo evidence and function. *Trends Cell Biol.*, **19** (8), 414–422.

39 Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., and Altman, S. (1983) The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, **35** (3), 849–857.

40 Kruger, K., Grabowski, P.J., Zaug, A.J., Sands, J., Gottschling, D.E., and Cech, T.R. (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. *Cell*, **31** (1), 147–157.

41 Nahvi, A., Sudarsan, N., Ebert, M.S., Zou, X., Brown, K.L., and Breaker, R.R. (2002) Genetic control by a metabolite binding mRNA. *Chem. Biol.*, **9** (9), 1043.

42 Winkler, W., Nahvi, A., and Breaker, R.R. (2002) Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature*, **419** (6910), 952–956.

43 Afonin, K.A., Bindewald, E., Yaghoubian, A.J., Voss, N., Jacovetty, E., Shapiro, B.A., and Jaeger, L. (2010) In vitro assembly of cubic RNA-based scaffolds designed in silico. *Nat. Nanotechnol.*, **5** (9), 676–682.

44 Chworos, A., Severcan, I., Koyfman, A.Y., Weinkam, P., Oroudjev, E., Hansma, H.G., and Jaeger, L. (2004) Building programmable Jigsaw puzzles with RNA. *Science*, **306** (5704), 2068–2072.

45 Dibrov, S.M., McLean, J., Parsons, J., and Hermann, T. (2011) Self-assembling RNA square. *Proc. Natl. Acad. Sci. U.S.A.*, **108** (16), 6405–6408.

46 Severcan, I., Geary, C., Chworos, A., Voss, N., Jacovetty, E., and Jaeger, L. (2010) A polyhedron made of tRNAs. *Nat. Chem.*, **2** (9), 772–779.

47 Ellington, A.D. and Szostak, J.W. (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature*, **346** (6287), 818–822.

48 Tuerk, C. and Gold, L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249** (4968), 505–510.

49 Walter, A.E., Turner, D.H., Kim, J., Lyttle, M.H., Müller, P., Mathews, D.H., and Zuker, M. (1994) Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc. Natl. Acad. Sci. U.S.A.*, **91** (20), 9218–9222.

50 Mathews, D.H., Sabina, J., Zuker, M., and Turner, D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288** (5), 911–940.

51 Isaacs, F.J., Dwyer, D.J., Ding, C., Pervouchine, D.D., Cantor, C.R., and Collins, J.J. (2004) Engineered riboregulators enable post-transcriptional control of gene expression. *Nat. Biotechnol.*, **22** (7), 841–847.

52 Bayer, T.S. and Smolke, C.D. (2005) Programmable ligand-controlled riboregulators of eukaryotic gene expression. *Nat. Biotechnol.*, **23** (3), 337–343.

53 Lou, C., Stanton, B., Chen, Y.-J., Munsky, B., and Voigt, C.A. (2012) Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nat. Biotechnol.*, **30** (11), 1137–1142.

54 Win, M.N. and Smolke, C.D. (2007) Targeted cleavage: tuneable cis-cleaving ribozymes. *Proc. Natl. Acad. Sci. U.S.A.*, **104** (38), 14881–14882.

55 Zadeh, J.N., Wolfe, B.R., and Pierce, N.A. (2010) Nucleic acid sequence design via efficient ensemble defect optimization. *J. Comput. Chem.*, **32** (3), 439–452.

56 Merino, E.J., Wilkinson, K.A., Coughlan, J.L., and Weeks, K.M. (2005) RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.*, **127** (12), 4223–4231.

57 Zuker, P.S.M. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9** (1), 133.

58 Dirks, R.M., Bois, J.S., Schaeffer, J.M., Winfree, E., and Pierce, N.A. (2007) Thermodynamic analysis of interacting nucleic acid strands. *SIAM Rev.*, **49** (1), 65–88.

59 Jaeger, L. and Leontis, N. (2000) Tecto-RNA: one-dimensional self-assembly through tertiary interactions *Angew. Chem. Int. Ed.*, **39** (14), 2521–2524.

60 Khisamutdinov, E.F., Jasinski, D.L., and Guo, P. (2014) RNA as a boiling-resistant anionic polymer material to build robust structures with defined shape and stoichiometry. *ACS Nano*, **8** (5), 4771–4781.

61 Wei, B., Dai, M., and Yin, P. (2012) Complex shapes self-assembled from single-stranded DNA tiles. *Nature*, **485** (7400), 623–626.

62 Ke, Y., Ong, L.L., Shih, W.M., and Yin, P. (2012) Three-dimensional structures self-assembled from DNA bricks. *Science*, **338** (6111), 1177–1183.

63 Myhrvold, C., Dai, M., Silver, P.A., and Yin, P. (2013) Isothermal self-assembly of complex DNA structures under diverse and biocompatible conditions. *Nano Lett.*, **13** (9), 4242–4248.

64 Win, M.N. and Smolke, C.D. (2007) A modular and extensible RNA-based gene-regulatory platform for engineering cellular function. *Proc. Natl. Acad. Sci. U.S.A.*, **104** (36), 14283–14288.

65 Beisel, C.L., Bayer, T.S., Hoff, K.G., and Smolke, C.D. (2008) Model-guided design of ligand-regulated RNAi for programmable control of gene expression. *Mol. Syst. Biol.*, **4** (224).

66 Qi, L., Lucks, J.B., Liu, C.C., Mutalik, V.K., and Arkin, A.P. (2012) Engineering naturally occurring trans-acting non-coding RNAs to sense molecular signals. *Nucleic Acids Res.*, **40** (12), 5775–5786.

67 Wilson, D.S. and Szostak, J.W. (1999) In vitro selection of functional nucleic acids. *Annu. Rev. Biochem.*, **68** (1), 611–647.

68 Lee, J.F., Hesselberth, J.R., Meyers, L.A., and Ellington, A.D. (2004) Aptamer database. *Nucleic Acids Res.*, **32** (Database issue), D95–D100.

69 Davis, J.H. and Szostak, J.W. (2002) Isolation of high-affinity GTP aptamers from partially structured RNA libraries. *Proc. Natl. Acad. Sci. U.S.A.*, **99** (18), 11616–11621.

70 Colas, P., Cohen, B., Jessen, T., Grishina, I., McCoy, J., and Brent, R. (1996) Genetic selection of peptide aptamers that recognize and inhibit cyclin-dependent kinase 2. *Nature*, **380** (6574), 548–550.

71 Shangguan, D., Li, Y., Tang, Z., Cao, Z.C., Chen, H.W., Mallikaratchy, P., Sefah, K., Yang, C.J., and Tan, W. (2006) Aptamers evolved from live cells as effective molecular probes for cancer study. *Proc. Natl. Acad. Sci. U.S.A.*, **103** (32), 11838–11843.

72 Hanes, J. and Plückthun, A. (1997) In vitro selection and evolution of functional proteins by using ribosome display. *Proc. Natl. Acad. Sci. U.S.A.*, **94** (10), 4937–4942.

73 Roberts, R.W. and Szostak, J.W. (1997) RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **94** (23), 12297–12302.

74 Bayer, T.S., Booth, L.N., Knudsen, S.M., and Ellington, A.D. (2005) Arginine-rich motifs present multiple interfaces for specific binding by RNA. *RNA*, **11** (12), 1848–1857.

75 Bertrand, E., Chartrand, P., Schaefer, M., Shenoy, S.M., Singer, R.H., and Long, R.M. (1998) Localization of ASH1 mRNA particles in living yeast. *Mol. Cell*, **2** (4), 437–445.

76 Brodsky, A.S. and Silver, P.A. (2000) Pre-mRNA processing factors are required for nuclear export. *RNA*, **6** (12), 1737–1749.

77 Fusco, D., Accornero, N., Lavoie, B., Shenoy, S.M., Blanchard, J.-M., Singer, R.H., and Bertrand, E. (2003) Single mRNA molecules demonstrate probabilistic movement in living mammalian cells. *Curr. Biol.*, **13** (2), 161–167.

78 Golding, I. and Cox, E.C. (2004) RNA dynamics in live *Escherichia coli* cells. *Proc. Natl. Acad. Sci. U.S.A.*, **101** (31), 11310–11315.

79 Valencia-Burton, M., McCullough, R.M., Cantor, C.R., and Broude, N.E. (2007) RNA visualization in live bacterial cells using fluorescent protein complementation. *Nat. Methods*, **282** (5387), 296–298.

80 Guo, S., Tschammer, N., Mohammed, S., and Guo, P. (2005) Specific delivery of therapeutic RNAs to cancer cells via the dimerization mechanism of phi29 motor pRNA. *Hum. Gene Ther.*, **16** (9), 1097–1109.

81 Ponchon, L. and Dardel, F. (2007) Recombinant RNA technology: the tRNA scaffold. *Nat. Methods*, **4** (7), 571–576.

82 Schifferer, M. and Griesbeck, O. (2009) Application of aptamers and autofluorescent proteins for RNA visualization. *Integr. Biol.*, **1** (8), 499–505.

83 Le, T.T., Harlepp, S., Guet, C.C., Dittmar, K., Emonet, T., Pan, T., and Cluzel, P. (2005) Real-time RNA profiling within a single bacterium. *Proc. Natl. Acad. Sci. U.S.A.*, **102** (2), 9160–9164.

84 Keiler, K.C. (2011) RNA localization in bacteria. *Curr. Opin. Microbiol.*, **14** (2), 155–159.

85 Broude, N.E. (2011) Analysis of RNA localization and metabolism in single live bacterial cells: achievements and challenges. *Mol. Microbiol.*, **80** (5), 1137–1147.

86 Ozawa, T., Natori, Y., Sato, M., and Umezawa, Y. (2007) Imaging dynamics of endogenous mitochondrial RNA in single living cells. *Nat. Methods*, **4** (5), 413–419.

87 Yiu, H.-W., Demidov, V.V., Toran, P., Cantor, C.R., and Broude, N.E. (2011) RNA detection in live bacterial cells using fluorescent protein complementation triggered by interaction of two RNA aptamers with two RNA-binding peptides. *Pharmaceuticals*, **4** (3), 494–508.

88 Valencia-Burton, M., Shah, A., Sutin, J., Borogovac, A., McCullough, R.M., Cantor, C.R., Meller, A., and Broude, N.E. (2009) Spatiotemporal patterns and transcription kinetics of induced RNA in single bacterial cells. *Proc. Natl. Acad. Sci. U.S.A.*, **106** (38), 16399–16404.

89 Agapakis, C.M., Boyle, P.M., and Silver, P.A. (2012) Natural strategies for the spatial optimization of metabolism in synthetic biology. *Nat. Chem. Biol.*, **8** (6), 527–535.

90 Chen, A.H. and Silver, P.A. (2012) Designing biological compartmentalization. *Trends Cell Biol.*, **22** (12), 662–670.

91 Erkelenz, M., Kuo, C.-H., and Niemeyer, C.M. (2011) DNA-mediated assembly of cytochrome P450 BM3 subdomains. *J. Am. Chem. Soc.*, **133** (40), 16111–16118.

92 Liu, M., Fu, J., Hejesen, C., Yang, Y., Woodbury, N.W., Gothelf, K., Liu, Y., and Yan, H. (2013) A DNA tweezer-actuated enzyme nanoreactor. *Nat. Commun.*, **4**, 2127.

93 Niemeyer, C.M., Koehler, J., and Wuerdemann, C. (2002) DNA-directed assembly of bienzymic complexes from in vivo biotinylated NAD (P) H: FMN oxidoreductase and luciferase. *ChemBioChem*, **3** (2), 242–245.

94 You, M., Wang, R.-W., Zhang, X., Chen, Y., Wang, K., Peng, L., and Tan, W. (2011) Photon-regulated DNA-enzymatic nanostructures by molecular assembly. *ACS Nano*, **5** (12), 10090–10095.

95 Conrado, R.J., Wu, G.C., Boock, J.T., Xu, H., Chen, S.Y., Lebar, T., Turnsek, J., Tomsic, N., Avbelj, M., Gaber, R., Koprivnjak, T., Mori, J., Glavnik, V., Vovk, I., Bencina, M., Hodnik, V., Anderluh, G., Dueber, J.E., Jerala, R., and DeLisa, M.P. (2012) DNA-guided assembly of biosynthetic pathways promotes improved catalytic efficiency. *Nucleic Acids Res.*, **40** (4), 1879–1889.

96 Moon, T.S., Dueber, J.E., Shiue, E., Prather, K.L.J., and Prather, K.L. (2010) Use of modular, synthetic scaffolds for improved production of glucaric acid in engineered *E. coli*. *Metab. Eng.*, **12** (3), 298–305.

97 Delebecque, C.J., Silver, P.A., and Lindner, A.B. (2012) Designing and using RNA scaffolds to assemble proteins in vivo. *Nat. Protoc.*, **7** (10), 1797–1807.

98 Ducat, D.C., Sachdeva, G., and Silver, P.A. (2011) Rewiring hydrogenase-dependent redox circuits in cyanobacteria. *Proc. Natl. Acad. Sci. U.S.A.*, **108** (10), 3941–3946.

99 Schirmer, A., Rude, M.A., Li, X., Popova, E., and del Cardayre, S.B. (2010) Microbial biosynthesis of alkanes. *Science*, **329** (5991), 559–562.

100 Torella, J.P., Ford, T.J., Kim, S.N., Chen, A.M., Way, J.C., and Silver, P.A. (2013) Tailored fatty acid synthesis via dynamic control of fatty acid elongation. *Proc. Natl. Acad. Sci. U.S.A.*, **110** (28), 11290–11295.

101 Barros, L.F. and Martínez, C. (2007) An enquiry into metabolite domains. *Biophys. J.*, **92** (11), 3878–3884.

102 Lee, H., DeLoache, W.C., and Dueber, J.E. (2012) Spatial organization of enzymes for metabolic engineering. *Metab. Eng.*, **14** (3), 242–251.

103 Fu, J., Liu, M., Liu, Y., Woodbury, N.W., and Yan, H. (2012) Interenzyme substrate diffusion for an enzyme cascade organized on spatially addressable DNA nanostructures. *J. Am. Chem. Soc.*, **134** (12), 5516–5519.

104 Ponchon, L., Catala, M., Seijo, B., El Khouri, M., Dardel, F., Nonin-Lecomte, S., and Tisne, C. (2013) Co-expression of RNA-protein complexes in *Escherichia coli* and applications to RNA biology. *Nucleic Acids Res.*, **41** (15), e150.

105 Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., and Smith, H.O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, **6** (5), 343–345.

106 Kosuri, S., Eroshenko, N., LeProust, E.M., Super, M., Way, J., Li, J.B., and Church, G.M. (2010) Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat. Biotechnol.*, **28** (12), 1295–1299.

107 Liang, J.C., Bloom, R.J., and Smolke, C.D. (2011) Engineering biological systems with synthetic RNA molecules. *Mol. Cell*, **43** (6), 915–926.

108 Wang, K., Neumann, H., Peak-Chew, S.Y., and Chin, J.W. (2007) Evolved orthogonal ribosomes enhance the efficiency of synthetic genetic code expansion. *Nat. Biotechnol.*, **25** (7), 770–777.

109 Neumann, H., Wang, K., Davis, L., Garcia-Alai, M., and Chin, J.W. (2010) Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome. *Nature*, **464** (7287), 441–444.

110 Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013) RNA-guided human genome engineering via Cas9. *Science*, **339** (6121), 823–826.

111 Sachdeva, G., Garg, A., Godding, D., Way, J., and Silver, P.A. (2014) In vivo co-localization of enzyme on RNA-scaffolds increases metabolic production in a geometrically dependent manner. *Nucleic Acids Res.* doi: 10.1093/nar/gku617

112 Douglas, S.M., Bachelet, I., and Church, G.M. (2012) A logic-gated nanorobot for targeted transport of molecular payloads. *Science*, **335** (6070), 831–834.

113 Fu, J. and Yan, H. (2012) Controlled drug release by a nanorobot. *Nat. Biotechnol.*, **30** (5), 407–408.

114 Robinson-Mosher, A., Shinar, T., Silver, P.A., and Way, J. (2013) Dynamics simulations for engineering macromolecular interactions. *Chaos*, **23** (2), 025110.

115 Wu, C.-H., Lockett, M.R., and Smith, L.M. (2012) RNA-mediated gene assembly from DNA arrays. *Angew. Chem. Int. Ed.*, **51** (19), 4628–4632.

116 Lucks, J.B., Mortimer, S.A., Trapnell, C., Luo, S., Aviran, S., Schroth, G.P., Pachter, L., Doudna, J.A., and Arkin, A.P. (2011) Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl. Acad. Sci. U.S.A.*, **108** (27), 11063–11068.

117 Shu, X., Lev-Ram, V., Deerinck, T.J., Qi, Y., Ramko, E.B., Davidson, M.W., Jin, Y., Ellisman, M.H., and Tsien, R.Y. (2011) A genetically encoded tag for correlated light and electron microscopy of intact cells, tissues, and organisms. *PLoS Biol.*, **9** (4), e1001041.

118 Martell, J.D., Deerinck, T.J., Sancak, Y., Poulos, T.L., Mootha, V.K., Sosinsky, G.E., Ellisman, M.H., and Ting, A.Y. (2012) Engineered ascorbate peroxidase as a genetically encoded reporter for electron microscopy. *Nat. Biotechnol.*, 1–9.

119 Choi, H.M.T., Beck, V.A., and Pierce, N.A. (2014) Next-generation in situ hybridization chain reaction: higher gain, lower cost, greater durability. *ACS Nano*, **8** (5), 4284–4294.

120 Jungmann, R., Avendaño, M.S., Woehrstein, J.B., Dai, M., Shih, W.M., and Yin, P. (2014) Multiplexed 3D cellular super-resolution imaging with DNA-PAINT and exchange-PAINT. *Nat. Methods*, **11** (3), 313–318.

14

Sequestered: Design and Construction of Synthetic Organelles

Thawatchai Chaijarasphong[1] *and David F. Savage*[2,3]

[1] *Mahidol University, Faculty of Science, Department of Biotechnology, Rama VI Rd., Bangkok 10400, Thailand*
[2] *University of California, Department of Molecular and Cell Biology, 2151 Berkeley Way, Berkeley, CA 94720, USA*
[3] *University of California, Department of Chemistry, 2151 Berkeley Way, Berkeley, CA 94720, USA*

14.1 Introduction

Spatial organization is a design principle of life. At the most basic level, compartmentalization defines the living contents within an organism from the nonliving extracellular milieu. Inside the cell, the sequestration of processes into distinct organelles and spaces is a common strategy for enabling competing pathways. Compartmentalization therefore allows for concurrent metabolic processes that are thermodynamically out of equilibrium with each other. The chemiosmotic proton-motive force is a classic example, in which protons are pumped from the matrix of the mitochondrion into the intermembrane space, using free energy derived from electron transfer [1]. By exquisitely regulating this gradient, the cell can capture its stored energy to synthesize adenosine triphosphate (ATP). Collapse of the gradient to equilibrium eliminates the mitochondria's ability to synthesize ATP and results in cell death.

From a biocatalysis point of view, compartmentalization creates a number of potential advantages for the engineer. First, it offers an additional way to regulate pathways [2]. Metabolites can be marked for specific processes in a regulated fashion, such as in the case of fatty acid oxidation and synthesis, which use orthogonal pools of fatty acyl-coenzyme A or fatty acyl–acyl carrier protein (ACP), respectively. Enzymes can also be selectively regulated via localization, such as in the glycosome, a peroxisome-derived organelle found in protozoa [3]. As its name suggests, the glycosome sequesters the first seven enzymes of glycolysis into a separate compartment. Its function appears to be regulatory. The glycosomal enzymes do not possess typical allosteric regulation (e.g., feedback inhibition of phosphofructokinase), and it is thought that compartmentalization achieves the same effect [4].

Synthetic Biology: Parts, Devices and Applications, First Edition. Edited by Christina Smolke.

Co-localization of a pathway also ensures substrate channeling of intermediates between enzymatic steps to improve both kinetics and yield and reduce host toxicity [5]. Channeling commonly occurs in multifunctional enzymes where a labile or toxic molecule is passed from one active site to another via a protein channel. Examples include tryptophan synthase (indole intermediate), acetyl-CoA synthase/carbon monoxide dehydrogenase (carbon monoxide), and carbamoyl phosphate synthase (ammonia) [6–8]. Similar mechanisms occur in bacterial microcompartments (BMCs), large proteinaceous shells that encapsulate short metabolic pathways. These will be discussed in greater detail later, but briefly, various evidence suggests these protein complexes are able to sequester/channel both volatile substrates (CO_2, acetaldehyde) and those potentially toxic (propanal) to the rest of the cell [9–11]. In a related context, it is important to note that self-assembly of enzymes and pathways into large complexes is more common than previously realized. It is perhaps an inevitable outcome of the fact that metabolic enzymes are highly expressed and allosterically regulated [12]. Thus, substrate channeling is often critical to metabolic pathway function.

Recent synthetic biological efforts have leveraged these principles for improved biocatalysis. The goal of metabolic engineering is to produce important chemicals, such as pharmaceuticals, materials, and biofuels, from cheap and sustainable biomass [13, 14]. Doing so requires high productivities and yields for engineered pathways, but this optimization is often counter to the growth and fitness of the host organism. Drawing inspiration from nature, one promising metabolic engineering strategy is to repurpose organelles or protein complexes as cellular factories for improving the performance of engineered pathways – in other words, to engineer synthetic organelles. In a striking example of this strategy, Dueber and colleagues have engineered scaffold proteins from the yeast mitogen-activating signaling cascade as an enzymatic assembly line to improve production titers of the isoprenoid precursor pathway nearly 80-fold while reducing intermediate toxicity [15]. Similarly, Sachdeva *et al.* improved synthesis of pentadecane from fatty acyl-ACP by co-localizing fatty acyl-ACP reductase and aldehyde-deformylating oxygenase to an RNA scaffold, providing a strategy for optimizing microbial biofuel production [16].

Building upon the idea of enhancing pathway flux through co-localization, various molecular chassis and metabolic engineering strategies have been developed to facilitate catalysis. To this end, here we review recent advances and open questions in the engineering and use of synthetic organelles for bioengineering applications (Figure 14.1). As most research heretofore has centered on metabolism, our focus is largely on metabolic engineering applications. The physical composition of an organelle – whether it is made from lipids or proteins – profoundly shapes potential uses, so our review is conceptually broken down into these two areas. Finally, it should be noted that the compartmentalization of engineering metabolism spans many orders of magnitude, from metabolically engineered cocultures of microbes down to single enzymes (Figure 14.1). We will focus on the middle regime, from protein compartments to repurposing existing organelles, but direct the engaged reader to previous reviews focused on the former [17] and latter [18, 19].

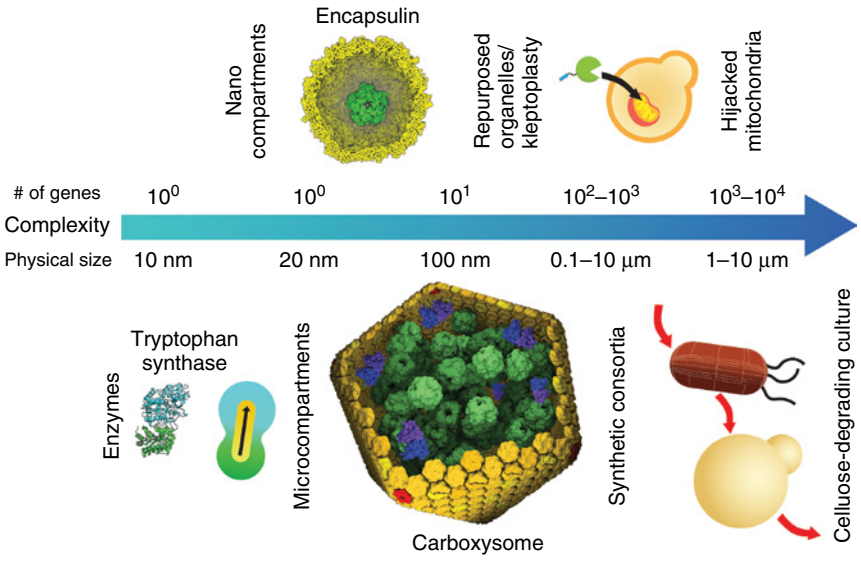


Figure 14.1 Possible strategies for engineering a synthetic organelle. Complexity of intra- and intercellular spatial organization spans from enzymes with inherent substrate channeling to symbiotic cocultures. This review highlights work in the middle ground, from nanocompartments to repurposed organelles.

14.2 On Organelles

Advances in imaging and comparative genomics have muddled the latter twentieth-century definition of an organelle as a specialized lipid-enclosed compartment found only in eukaryotes [20, 21]. It is now clear that many prokaryotes contain topologically distinct membrane compartments and that proteinaceous BMCs also found in prokaryotes possess metabolic features similar to complex structures such as mitochondria [9, 22]. In the early years of light microscopy, beginning with Möbius in the late 1800s, many cytoplasmic features including ribosomes, flagella, and the centriole were labeled with the diminutive organelle. Given recent results and historically ambiguity, we therefore propose to adapt a more relaxed definition in the context of this review: an organelle is simply a physically delimited compartment within the cell.

An alternative viewpoint, particularly for the synthetic biologist, is to ask what is required to repurpose an existing organelle or construct one *de novo*. In this light, four important intertwined, but distinct, themes emerge (Figure 14.2). The first is *targeting*. To accomplish orthogonal function in a specific compartment, it is essential to have selective targeting of biochemical activities (i.e., typically enzymes). Nature widely leverages the specificity inherent to protein–protein interactions through the use of signal sequences. Engineering an organelle requires extensive knowledge of targeting, specificity of this process, and ideally, how the stoichiometry of targeted components can be adjusted to control activity of individual proteins.

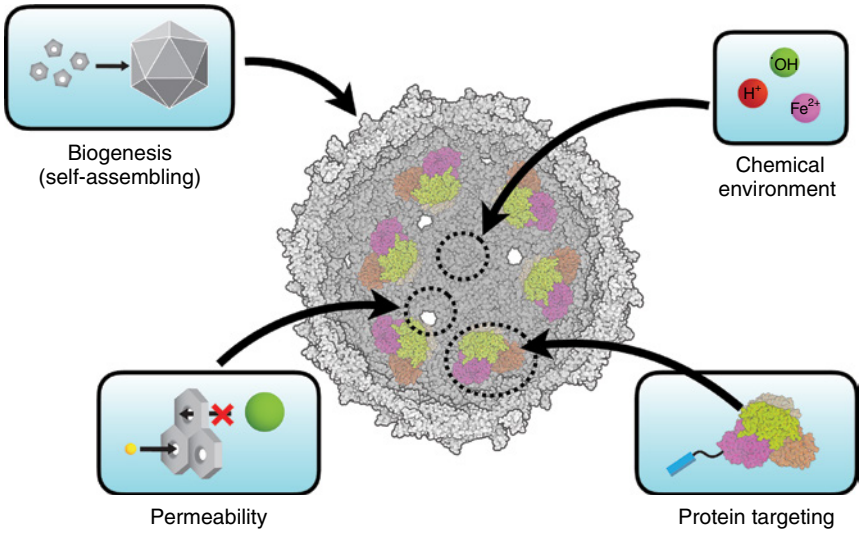


Figure 14.2 Four core design principles for a synthetic organelle.

Catalysis is often the motivating factor for organelle engineering, leading to two linked properties – compartment *permeability* and its inherent *chemical environment*. Permeability is the selectivity of the surrounding membrane or protein shell that directly affects what can diffuse across or be transported in and out of the compartment. In lipid-based organelles, selectivity is modulated by the nature of the membrane lipid content and types and specificities of integral membrane transporters or channels. In proteinaceous organelles, it is simply a function of the shell's diffusive permeability. In a related context, the chemical environment, set up by the interplay of both permeability and combined enzymatic activity taking place within the compartment, will control the concentrations of potential substrates and products, as well as general properties such as pH [23]. These concentrations will directly control both the thermodynamic equilibrium of a particular process and its kinetics, profoundly shaping the catalytic potential of an organelle.

Finally, it is important to have a working understanding of organelle *biogenesis*. Biogenesis is the process of organelle self-organizing and will control organelle shape, size, and copy number [24]. Repurposing efforts focused on existing lipid-based organelles have so far shied away from extensive remodeling, but it is logical to assume future efforts will enable the complete refactoring of existing structures or even the *de novo* creation of novel compartments. Engineered biogenesis has had more success in the protein-based space, as the genetic information required for synthesis is far less. BMCs contain roughly 10–15 proteins, and there are established systems for the transgenic expression of microcompartments in new organismal hosts.

In understanding these four properties, we therefore seek a deep understanding of organelle structure and function. Put another way, the synthetic biology-minded goal of engineering novel organelles represents hypothesis testing to an

extreme – that is, if you can't build it, you don't understand it. From a historical perspective, it is important to realize that these are not new ideas. We simply have better molecular tools. To end, we quote from F.H. Gaertner, who posited similar ideas over three decades ago:

> The degree to which the majority of the cytosolic enzyme systems may be organized, and the manner in which such organization would endow these systems with one or more of the unique catalytic properties, stand as open questions. In order to answer these questions fully, our ultimate challenge may be to take a cell apart and put it back together again. [25]

14.3 Protein-Based Organelles

We begin with protein complexes, which represent a modular route to synthetic organelle construction. Nature widely uses substrate channeling in enzymatic complexes [5, 26], but decoupling compartmentalization from inherent enzymatic function is challenging. For example, the substrate-channeling tunnel of tryptophan synthase is structurally intertwined with the α and β subunits and their active sites. Altering enzymatic function while maintaining channeling between active sites would require a tremendous protein engineering effort. A more sensible starting point is therefore an *a priori* functionally decoupled system, in which compartmentalization is a property distinct from enzymatic function.

14.3.1 Bacterial Microcompartments

BMCs are proteinaceous organelles that are functionally decoupled into shell proteins and cargo proteins (Figure 14.3a) [9, 27, 28]. The cargo proteins possess enzymatic function and generally constitute a small metabolic pathway of two to four reactions. The widespread prevalence and rich diversity of these organelles became evident in a recent bioinformatics study, which identified 23 types of BMCs in 23 phyla of bacteria [29–31, 134]. Functionally, BMCs can be grouped into two main categories: anabolic and catabolic microcompartments [11, 29]. The only known member of the anabolic group is the carboxysome, which performs carbon dioxide fixation in photoautotrophic and chemotrophic bacteria [32]. Catabolic BMCs (also called metabolosomes), as the name suggests, perform various catabolic reactions that help break down nutrients. This class of BMCs accounts for most of the diversity reported [17, 29], but only two members have been extensively characterized: propanediol-utilizing (PDU) microcompartment [135, 136] and ethanolamine-utilizing (EUT) microcompartment [33, 137].

Despite this divergence, the three most-studied BMCs – carboxysome, PDU, and EUT – share similar structural arrangement and mode of function (Figure 14.3a). The shell is formed principally by a ~100-amino-acid α/β protein possessing a canonical BMC domain (Pfam00936), which oligomerizes into a homohexamer roughly 70 Å in diameter [34]. Subsequently, this hexamer self-assembles into larger sheetlike structures that form the facets of the BMC shell.

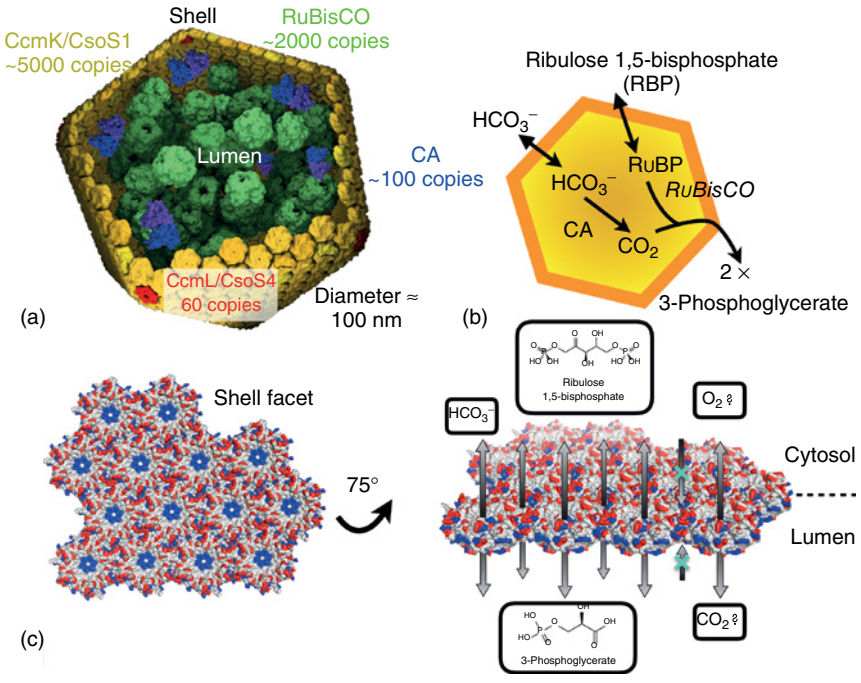


Figure 14.3 The structure and function of a bacterial microcompartment, the carboxysome. (a) Structural model showing some of the structural components and enzymatic cargo. (b) Schematic of the carbon-concentrating mechanism of carboxysome function. (c) Model of shell permeability based on CcmK4 (PDB: 2A10).

Examples of these proteins are CcmK2 and CsoS1A from the CB, PduA from the PDU, and EutM from the EUT [35]. There are roughly 4000–5000 copies of the protein per shell. Additionally, there is a minor protein component of ~90 amino acids that does not contain the canonical BMC domain and instead oligomerizes into a pentamer (Pfam03319), as determined via X-ray crystallography [36]. Cryo-electron microscopy (EM) studies of purified BMCs suggest that the overall structure possesses a roughly icosahedral form [37, 38]. Topological constraints therefore dictate that the pentameric protein forms the vertices of the icosahedron. There are 12 vertices in an icosahedron, placing the exact stoichiometry of the monomer at 60 copies. Biochemical evidence also suggests the pentamer is of very low abundance in purified BMCs. Examples of this family include CcmL and CsoS4A from the CB, PduN in the PDU, and EutN from the EUT. Intriguingly, although EutN crystallizes as a hexamer, protease cleavage experiments suggest it is a pentamer in solution [36, 39]. This heterogeneity in quaternary structure may explain the somewhat irregular form of isolated PDUs and EUTs in comparison with the more icosahedral-like CBs. The overall polyhedral structure therefore consists of (roughly) 20 triangular facets built from thousands of shell proteins. Depending on the BMC and preparation protocol, the overall structure is of size 80–400 nm. Finally, cargo proteins are targeted to the lumen of the BMC through protein–protein interactions with the inside face

of the shell. Depending on the BMC, there are thousands of copies of protomers (i.e., ~2000 RuBisCO (ribulose 1,5-bisphosphate carboxylase/oxygenase) monomers in a CB) targeted to the lumen [37]. The specific mechanisms of targeting are discussed in greater detail later.

Before delving into the prospects of reengineering BMCs, it is important to understand their function in the native context. The CB was the first BMC to be discovered and characterized, and it remains the paradigm for BMC function [40]. We give an overview of CB function here to highlight themes of BMC function (Figure 14.3b). Organisms that assimilate carbon using the Calvin–Benson cycle must compensate for the low affinity of RuBisCO for CO_2 and for its promiscuity – RuBisCO can also fix O_2 in the same reaction at a cost to the cell. To overcome these limitations, cyanobacteria and many chemoautotrophs employ a carbon-concentrating mechanism, which consists of inorganic carbon transporters to increase intracellular bicarbonate levels, and the CB to facilitate carbon fixation [41, 42]. After bicarbonate is actively transported into the cell, it passively crosses the CB shell (details on this later in text) and enters the CB lumen. The CB encapsulates two enzymes, carbon anhydrase and RuBisCO. Carbonic anhydrase interconverts bicarbonate into CO_2 and OH^-, and RuBisCO fixes this CO_2 onto ribulose 1,5-bisphosphate, which must also enter the lumen, and produces two molecules of 3-phosphoglycerate. 3-Phosphoglycerate then diffuses out of the CB and enters the reductive phase of the Calvin–Benson cycle. Although modeling indicates the major mechanism benefiting the reaction is an increased local concentration of CO_2 to improve the catalytic rate [23], additional possible mechanisms include excluding the competing substrate O_2 from the lumen, improving CO_2 channeling from carbonic anhydrase to RuBisCO via tight clustering of the enzymes [43], and raising the local pH around RuBisCO to increase its catalytic activity (Figure 14.3c). Most of the experimental evidence for these hypotheses is indirect, for example, catalytically dead carbonic anhydrase mutants require high CO_2 concentrations to grow [30], suggesting further physiological experiments will be needed to describe the actual mechanism(s) used by the CB to facilitate carbon fixation. Finally, it is important to note that CB comes in two forms, the so-called α and β type [44]. They are differentiated by sequence in their shell and cargo proteins, particularly carbonic anhydrase, and by their genomic organization. In general, genes for α-CBs occur together in a single operon in the genome, while the β-CB regulon is composed of genes spread across the genome. Despite these evolutionary differences, their catalytic activity is the same and their physiological role is assumed to be similar [45].

14.3.1.1 Targeting

Although the shell is the defining feature of BMCs, it is the targeting of cargo that endows function. Targeting is mediated via protein–protein interactions and probably occurs concurrently with assembly of the shell itself. The first direct evidence for a shell-interacting motif was found in the PDU [46]. In this case, bioinformatic analysis of the propionaldehyde dehydrogenase (PduP) reveals an N-terminal extension found only in PDU-containing organisms. Alanine scanning, among other biochemical experiments, has shown this putative α-helical

signal sequence can interact with the hexameric shell proteins (PduA, PduJ, and PduK), and inclusion of this sequence allows encapsulation of foreign cargo [47, 48]. In addition, subsequent studies also reveal that the short N-terminal extension of the medium subunit (PduD) of adenosylcobalamin diol dehydratase (PduCDE) and an unknown protein PduV can target their respective cargo to PDU [49, 50]. Besides PDU, other catabolic BMCs, including EUT and a glycyl radical-based propanediol utilization (GRP) microcompartment, use signal peptides to encapsulate their respective cargo [51–53]. Strikingly, these targeting peptides have been shown to enable targeting of green fluorescent protein (GFP) to PDU, suggesting that the targeting specificity for these BMCs is not stringent and may be determined by the composition rather than the sequence of the targeting peptides [53]. This relaxed specificity allowed for the *de novo* construction of synthetic signal peptide for PDU targeting. With the growing repertoire of natural and synthetic signaling peptides, it may soon be possible to encapsulate multiple enzymes in a BMC to constitute a longer metabolic pathway. While the mechanism of encapsulation via signaling peptide is not completely understood, interesting applications have already emerged, including the construction of an ethanol nanoreactor by encapsulating pyruvate decarboxylase and alcohol dehydrogenase in PDU [47] and compartmentalization of polyphosphate kinase (PPK1) in PDU to enhance the conversion of biological phosphates to cellular polyphosphate [54].

In contrast to catabolic BMCs, the targeting strategy used by carboxysomes is less well characterized, and most of the understanding came from β-CBs. Pull-down and yeast two hybrid experiments probing components of the cyanobacterial β-CB revealed that RuBisCO is anchored to the shell via specific interactions with the protein CcmM, which acts as an intermediate bridge between enzymatic cargo and the shell. CcmM also possesses a nonfunctional carbonic anhydrase-like domain and recruits the functional carbonic anhydrase, CcaA, forming a functional carbon-fixing complex [37, 38]. More recently, a CB protein CcmN was found to be essential for the shell recruitment during carboxysome assembly, and its deletion resulted in a large shell-less RuBisCO aggregate [55]. A C-terminal extension of this protein appears to interact with the major shell hexamer CcmK2 and is sufficient for targeting the GFP into CBs [56]. Therefore, CcmN may be the actual mediator between the shell and the CcmM/CcaA/RuBisCO complex discussed earlier. In α-CBs, homologs of CcmM and CcmN are not present, but a poorly characterized protein called CsoS2 may perform an analogous function. CsoS2 is an intrinsically disordered protein with many amino acid repeats [57, 58]. These properties are often associated with proteins that function as "assembly coordinators" for large complexes, thus providing an important clue about the function of CsoS2 [59]. As evidence of the necessity of CsoS2 to α-CB assembly, deletion of CsoS2 in *Halothiobacillus neapolitanus* abolishes carboxysome formation and renders the organism high-CO_2 requiring (HCR) [57]. Interestingly, it was shown that one *csoS2* coding sequence produces two CsoS2 isoforms via a co-translational mechanism [58], reminiscent of CcmM and CcmN in β-CBs. If the CsoS2 isoforms are indeed functionally analogous to CcmM and CcmN, understanding the way they interact with other carboxysomal proteins may shed light on how cargo in α-CB is encapsulated. Additional

work is required to understand the true role of CsoS2 and to develop a strategy to target foreign cargo to α-CBs.

Despite these advances, several outstanding questions in carboxysome targeting remain. Firstly, the stoichiometry of cargo proteins can vary over 10-fold, that is, there are roughly 100 protomers of carbonic anhydrase and 2000 protomers of RuBisCO, yet there is no mechanistic explanation for how this can be programmed through protein–protein interactions alone. This will be critical to understand as future engineers attempt to balance flux through multistep enzymatic pathways. Secondly, little is known about protein targeting in the α-CB. The α-CB from *H. neapolitanus* is a structurally robust BMC that can assemble without cargo and be transgenically expressed in *Escherichia coli*, making it an intriguing chassis for synthetic biological purposes [60, 61]. Making this a reality, however, will ultimately require a complete biophysical understanding of the targeting motifs and mechanisms.

14.3.1.2 Permeability

The structure of the shell proteins is thought to control permeability of the BMC (Figure 14.3c). X-ray crystal structures of various hexameric and pentameric shell proteins have revealed pores along the major axis of symmetry that, in principle, would facilitate passive diffusion of substrates and products. The pores are generally small (4–6 Å in diameter), implying specificity [62]. In the case of the CB, positively charged residues are found at the narrowest area of the pore, suggesting a mechanism for screening for negatively charged molecules, such as the substrates/products bicarbonate, ribulose 1,5-bisphosphate, and 3-phosphoglycerate, and against molecules without a dipole such as O_2. Although there are few permeability measurements to support these hypotheses, physiological data clearly implies that there is minimal photorespiration (fixation of O_2) when RuBisCO is inside the CB, suggesting O_2 exclusion may be one effect of encapsulating RuBisCO [63, 64]. In addition, *csoS4*-disrupted *H. neapolitanus* have a HCR phenotype, and their CBs leak CO_2 as interpreted by kinetic experiments [65]. Thus, a tight BMC shell appears to act as a gas barrier to exclude O_2 and sequester CO_2. This theme is seen in other BMCs, as well. For instance, *Salmonella enterica* mutants that cannot produce PDU accumulate 10-fold increased levels of propanal in the cytosol [11]. Aldehydes, as nonspecific crosslinkers, damage DNA, and the 10× increase in propanal levels proved to be highly mutagenic. Similarly, alteration of pore-lining residues in PduA resulted in propanal leakage, reduced 1,2-propanediol influx, and increased glycerol influx, further substantiating the role of PDU shell as a selective diffusion barrier [66]. In the case of the EUT, shell mutants also leak their intermediate, acetaldehyde, but here, physiological data supports the hypothesis that the sequestration acts to stop the loss of a volatile intermediate out of the pathway. Thus, BMC shells can achieve many catalytic goals – enhancing pathway specificity and yield while reducing toxicity – by tuning their permeability.

Recent X-ray structures highlight an expanded toolkit for altering shell permeability. The relatively small size of most pores (4–6 Å) is at odds with the required permeability for larger substrates such as ribulose 1,5-bisphosphate or the cofactors coenzyme A and NAD used in the PDU and EUT. Kerfeld and colleagues

have solved the structure of the CsoS1D protein from the α-CB, which possesses a tandem BMC domain and forms a homotrimeric pseudohexamer with a much larger pore of 14 Å [67]. Although this protein is of low abundance and only recently detected in purified CBs, it may play an important role in allowing larger molecules passage into and out of the CB [61, 68]. Even more intriguingly, CsoS1D also crystallized in alternate conformations with both an open and closed pore. In follow-up work, Kerfeld *et al.* solved the structure of the orthologous tandem repeat protein from the β-CB and again observed open and closed forms [69]. More recently, EutL has been shown to have negative allosteric regulation for pore opening by ethanolamine, and disulfide bonding may play a role in modulating the binding affinity toward ethanolamine [70]. These findings raise the possibility of posttranslational regulation of BMC permeability.

Catalyzing redox reactions is a critical component of PDU and EUT activity and recent results also highlight the role of the shell in these processes. Overexpression of the *Citrobacter freundii* PDU and its components in *E. coli* led to the surprising realization that the shell protein PduT contains an Fe–S cluster on its major symmetry axis [71, 72]. This was confirmed via electron paramagnetic resonance and X-ray crystallography. The midpoint potential was measured at +0.099 V, suggesting the cluster may help recycle NADH, produced during the oxidation of propionaldehyde to propionyl-CoA, back to NAD^+. Similarly, the shell protein GrpU of GRP microcompartment also coordinates Fe–S cluster [73]. While further experimental validation is required to demonstrate that these shells can truly participate in a redox reaction, it does underscore the potential for catalytic flexibility among BMCs.

14.3.1.3 Chemical Environment

A related property is chemical environment, including redox state, pH, intermediate concentrations, and cofactor status. This results from the interplay of shell permeability and enzymatic activity in the lumen, creating steady-state concentrations of molecular species different than what exists in the cytosol. For example, a recent mathematical model predicts that a relatively acidic carboxysome will exhibit higher equilibrium CO_2 concentration and, in turn, a higher degree of RuBisCO saturation [23]. This finding raises the possibility that the actual carboxysome may similarly be acidic in order to achieve maximum catalytic efficiency. Preliminary biochemical analyses of carbonic anhydrases from some β-CBs also suggest that the lumen environment may be oxidative, promote disulfide bond formation, and be a means of controlling protein activity [74, 75, 138]. Interestingly, CsoS2, the putative scaffolding protein of the α-CB, contains many cysteine residues, half of which are conserved across amino acid repeats. The abundance of cysteines may imply CsoS2's participation in disulfide-bonding network within the carboxysomal lumen, which, if true, would explain the exceptional robustness of α-CB.

The chemical environment is likely more extreme in the PDU and EUT. As described earlier, one explanation for PDU/EUT function is to sequester the buildup of toxic aldehyde intermediates away from the cytoplasm. *S. enterica* mutants with disrupted PDU shells accumulate 10× higher levels of cytosolic propionaldehyde (~15 mM), suggesting that luminal aldehyde concentrations are

extremely high [11]. Interestingly, there is little information on the effect such harsh conditions have on protein activity within the BMCs. Many metabolic pathway intermediates, such as the aldehydes present in the PDU, EUT, and many candidate biofuel pathways, will inevitably cause protein misfolding and inactivate individual enzymes in the complex. A key open biological question is therefore how the proteostasis of BMCs and their enzymatic content is regulated. A single BMC is on the order of 0.2% of total cellular protein (estimated from CB mass of ~250 MDa [75] and *E. coli* protein content from BioNumber 104879 [76]) and represents a tremendous investment for the cell. It remains to be seen whether BMCs are surveilled via the cell's proteostatic chaperones and, if so, whether this entire "costly" complex is turned over at the level of single inactive subunits or all at once.

14.3.1.4 Biogenesis

The critical information for biogenesis is an understanding of the genes and expression levels that are necessary and sufficient for BMC self-assembly and function. Early genetic, cloning, and sequencing efforts revealed that BMCs genes are often co-localized together in operons but that the degree of co-localization varies with each BMC. For example, α-CB genes cluster together as a single regulon in the genome of *H. neapolitanus*, while the β-CB regulon found in many cyanobacterial strains is composed of five operons spread across the genome (Figure 14.4). For this reason, successful efforts at reconstituting and transgenically expressing fully functional BMCs in a heterologous host have focused on those where genes are co-localized in a single genomic island. For example, screening of a *C. freundii* genomic DNA library identified a cosmid capable of endowing *E. coli* with the ability to metabolize 1,2-propanediol [71]. Sequencing of this cosmid and further molecular biology to narrow down the candidate genes identified a minimal subset important for the heterologous production of PDUs [50]. This has proven an important tool for studying the role of each gene in defining PDU structure and for identifying signal sequences. A similar approach was successful for the α-CB and EUT. Expression of the

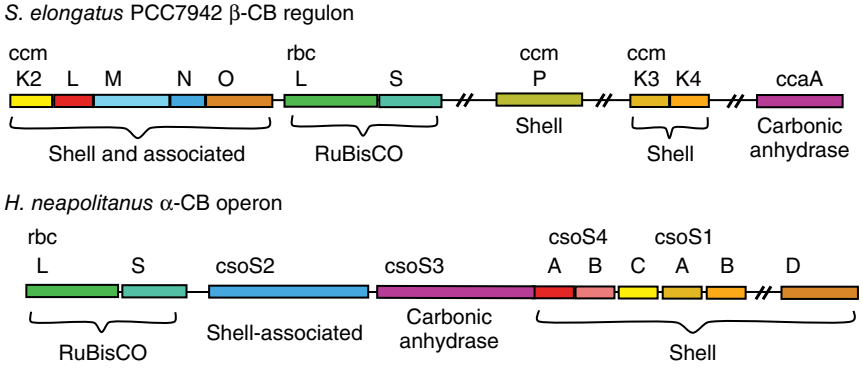


Figure 14.4 Genetic organization of α- and β-carboxysome regulons from two different bacteria. Color scheme indicates relatedness.

10 genes from *H. neapolitanus* led to fully formed CBs with a morphology nearly identical to those from the native host in *E. coli* [61]. Likewise, expression of 17 genes from *S. enterica* can also produce EUTs in *E. coli* [51, 137].

Although these early attempts have defined genetic sufficiency, more work will be required to define essential, or necessary, elements. The ultimate goal, of course, would be to have a minimal system for expressing empty protein shells and targeting novel proteins to the lumen. One open question is the role of each shell protein and how many different genes are required to synthesize well-formed polyhedral shells. Most regulons possess multiple copies of genes for the hexameric and pentameric protomers. Whether this is a gene dosage mechanism for high protein expression or there is a functional difference between paralogs remains to be seen. It should be noted that the function of shell paralogs may be determined by their genomic position. This issue was brought to attention by Chowdhury and colleagues, who demonstrated that PduJ is permeable to 1,2-propanediol only when it is expressed from the *pduA* locus [77]. It is unclear why such a location effect exists, although it is thought that nascent PduJ translated from different genomic regions may encounter different sets of binding partners. Following from this observation, it may be possible to alter permeability of a shell protein by changing its gene location in lieu of the labor-intensive site-directed mutagenesis.

Another factor related to biogenesis that will need to be clarified is the inherent stability of BMCs. It is known that CBs have a more icosahedral shape [78] and that α-CBs, in particular, are robust enough to be isolated from cells in a near-pristine form. Likewise, transgenic α-CBs display a somewhat native-like structure, suggesting that either the transcriptional and translational regulation of the operon sequence "ports" over better in *E. coli* or that the protein–protein interactions of α-CB self-assembly are inherently more robust. Future work will be necessary to clarify to what extent this hypothesis is true and whether it holds if BMCs are transgenically expressed in higher organisms such as yeast and plants. In fact, Lin and colleagues have already made the first attempt to produce carboxysomes in plants by expressing CcmM, CcmN, and three shell proteins (CcmK2, CcmL, and CcmO) from the β-CB in chloroplasts of *Nicotiana benthamiana*, but the resulting empty compartments were irregularly shaped [79]. It would be of special interest to determine whether a similar experiment with an α-CB would result in more morphologically normal particles.

14.3.2 Alternative Protein Organelles: A Minimal System

There are also several other self-assembling protein complexes that, in principle, could be adapted to function as an organelle. These include viral particles, large enzyme complexes such as lumazine synthase, the ribonucleoprotein vault complex, and the icosahedral encapsulin complex, all of which have been studied to some extent in attempts to engineer novel materials for both *in vivo* and *ex vivo* applications [80–83]. Since it is not possible to cover all such applications in an appreciable depth here, we refer the reader to previous reviews [84, 85] and instead focus on one particular complex – encapsulin – that is a minimal alternative to the more complex BMCs discussed earlier (Figure 14.5).

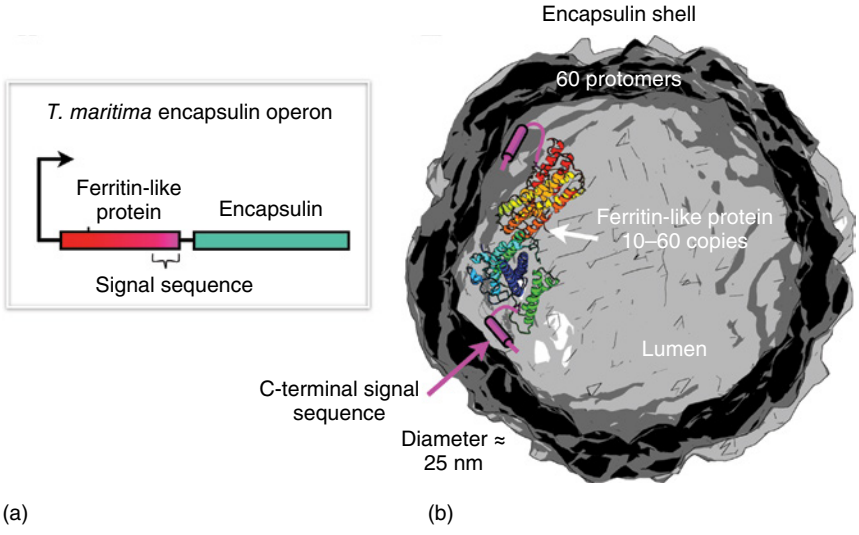


Figure 14.5 Model of an encapsulin. (a) Genomic organization in *T. maritima* highlight signal sequence of cargo protein. (b) Structural model based on encapsulin X-ray structure (PDB: 3DKT) and ferritin-like protein (PDB: 3HL1).

Encapsulins (also called nanocompartments) are a family of poorly characterized proteins that have the defining feature of assembling into 20–30 nm icosahedral complexes (Figure 14.5). The founding member, Linocin M18 from *Brevibacterium linens*, was discovered as a secreted protein with bactericidal activity, but recent results question this biological function [83, 86]. Since then, the number of predicted encapsulins has increased dramatically, with the latest bioinformatics study reporting over 900 putative encapsulins across 15 bacterial and 2 archaeal phyla [87]. Encapsulins, like BMCs, also appear to be diverse, with four families of capsids and seven classes of associated cargo [88]. Despite the diversity, only a small number of encapsulins have been biochemically characterized, including those from *Thermotoga maritima*, *Pyrococcus furiosus*, *Mycobacterium tuberculosis*, *Myxococcus xanthus*, and *Rhodococcus jostii* RHA1.

The X-ray crystal structures of three different encapsulins from *P. furiosus*, *T. maritima*, and most recently *M. xanthus* have been determined, clarifying many open structural and functional questions [83, 89] (Figure 14.5b). The structural shell is formed from a single protomeric protein that self-assembles into an icosahedral shell about 2 nm thick. In the *Pyrococcus* and *Myxococcus* variants, 180 protomers assemble into a structure 30 nm in diameter, while in the *Thermotoga* variant 60 protomers form a 20 nm structure, suggesting significant structural heterogeneity can exist between encapsulins. Like in BMCs, there are pores parallel to protomer symmetry axes. There are three distinct classes of pores, each possessing a diameter of about 5 Å, located at the interface between two adjacent protomers, sites of fivefold symmetry and sites of threefold symmetry. While the first two classes show interspecies conservation of the chemical property of the pore-lining residues, the same is not true for the

threefold axis pores – they are positively charged in the *Thermotoga* encapsulin while uncharged in many other classes. The explanation for this divergence is yet unknown, although it may reflect the nature of the small molecules that must traverse the shell.

This study also led to the identification of a putative signal sequence for encapsulins. Bioinformatic analysis has revealed that two classes of enzymes, peroxidases and ferritin-like proteins, preferentially cluster in minimal operons adjacent to the shell-forming encapsulin gene. Serendipitously, there was additional electron density in the *Thermotoga* structure abutting the inner face of the encapsulin shell. This density was of sufficient signal to identify a primary peptide sequence, which matched the C-terminus of the adjacent ferritin-like gene in the operon, establishing the link between the gene cluster and protein structure. Deletion of the C-terminal region also disrupted targeting of the enzyme to the lumen, confirming this sequence is essential for targeting [83]. By employing this targeting sequence, many studies have reported successful targeting of heterologous cargo into encapsulins [90–92].

Encapsulins therefore have many advantages as potential synthetic organelles. They are in many ways a minimal version of BMCs. They assemble from a single shell protein into a compartment possessing about 1/100 the volume. This genetic simplicity likely ensures porting structures between organisms will be easier than for BMCs (Figure 14.5a). Preliminary experiments agree with this hypothesis – encapsulins from many organisms including *B. linens*, *T. maritima*, *M. xanthus*, and *M. tuberculosis* can be expressed heterologously in *E. coli* [87]. As an additional advantage, encapsulins commonly display exceptional resistance to temperature, pH, denaturant, proteases, and mechanical compression [83, 90, 91, 93, 94]. Therefore, they may serve as appealing alternatives for applications that demand extreme conditions incompatible with other biological compartments. However, it appears the "addressability" – the number of proteins that can be targeted to its lumen – will be limited to one or two. For example, Snijder and colleagues employed native mass spectroscopy to show that one *B. linens* microcompartment precisely packages one hexamer of peroxidases, suggesting that the limited capacity is likely a valid concern [94]. In this vein, we imagine that very short pathways, that is, two steps with a single toxic intermediate, would make excellent candidates for encapsulation. Future work will also be required to understand and engineer shell permeability.

14.4 Lipid-Based Organelles

The alternative to protein-based complexes is to leverage the natural organization of metabolism found in eukaryotes – membranous organelles. This makes practical sense as many key pathways of catabolism and anabolism are segregated at the organelle level, as discussed in the following text. In addition, much of our biological understanding of these processes comes from the yeast *Saccharomyces cerevisiae*, which is arguably the most important organism for metabolic engineering (Figure 14.6). Thus, there is already a working understanding of organelle targeting, permeability, chemical environment, and biogenesis.

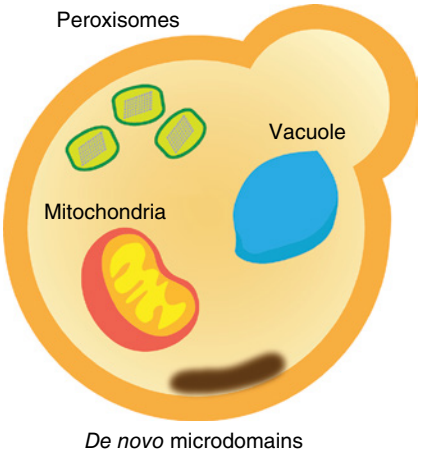


Figure 14.6 Schematic of potential synthetic organelles in the budding yeast.

14.4.1 Repurposing Existing Organelles

14.4.1.1 The Mitochondrion

The mitochondrion is the site of oxidative metabolism within the eukaryotic cell, facilitating both the citric acid cycle and β-oxidation of fatty acids, and is involved in numerous critical cell processes, including apoptosis [95]. Its function revolves around metabolism, and it possesses a singular chemical environment – relatively high pH (~8), low oxygen concentration, and a reducing redox environment [96]. Besides its commonly associated pathways, the mitochondrion also assists in several other biosynthetic pathways including iron–sulfur cluster biogenesis, heme biosynthesis, and, surprisingly, type II fatty acid synthesis [95, 97]. An understanding of mitochondrial biogenesis is still a work in progress, but its central role in metabolic diseases has led to a new appreciation of mitochondrial biology. Proteomics has revealed the parts list of mitochondrial components and putative pathways and also led to a deeper understanding of relevant synthetic biological issues such as protein targeting [98].

Recently, the mitochondria's unique catalytic potential has been leveraged for metabolic engineering approaches. Farnesyl diphosphate (FDP) is a 15-carbon metabolic intermediate in the isoprenoid pathway. It is synthesized from the two isomers: isopentenyl pyrophosphate (IPP) and dimethylallyl pyrophosphate (DMAPP). Once synthesized, FDP can be processed by so-called sesquiterpene synthases into numerous products including molecules that are potential biofuels and pharmaceuticals [99]. Farhi and colleagues hypothesized that since FDP stands at the intersection of isoprenoid biosynthesis, compartmentalization of its terminal reactions may enhance production [100]. This hypothesis was correct, and targeting of a sesquiterpene synthase to the mitochondria using the known N-terminal targeting sequence of *COX4* [101] led to a 3× increase in the final product, valencene [100]. Mitochondrial targeting of additional steps to produce the key intermediate, FDP, led to an additional twofold increase in final titers. Interestingly, despite these results, it is not known whether targeting is

successful due to higher levels of FDP in the mitochondrion or whether sesquiterpene synthase is simply more active in the mitochondrial matrix. Circumstantial evidence from plants suggests that the mitochondria matrix contains both IPP and DMAPP, agreeing with the former and alluding to a more complicated biosynthetic picture of mitochondrial function [102].

This approach has also been used for the production of another class of important chemicals, higher alcohols, which are potential gasoline replacements. Higher alcohols (aka fusel alcohols), such as isobutanol, are biosynthetically produced from the catabolism of amino acids via the Ehrlich pathway [103]. Interestingly, while the initial biosynthesis of amino acids occurs in mitochondria, the final Ehrlich decarboxylation and dehydrogenase reactions occur in the cytosol. Avalos and colleagues hypothesized that co-localizing the entire isobutanol pathway (derived from leucine) together could result in improved flux between enzymes and increased production [104]. This was indeed the case, and a complete mitochondrial-localized pathway resulted in a 260% increase in production. Interestingly, control experiments found that co-localization of the same enzymes to the cytoplasm improves yields only a 10% increase, suggesting the mitochondria possesses an inherent biosynthetic capability. Indeed, the mitochondrial targeting system has been used to optimize the production of acetoin [105] and fumarate [106] in yeasts and artemisinin in plant [107].

14.4.1.2 The Vacuole

The vacuole is the central degradative structure in fungi, such as *S. cerevisiae*, and is roughly the functional equivalent of the lysosome in mammals. It maintains a low pH and possesses numerous hydrolytic enzymes involved in catabolic processes [108]. These properties led to the classic notion of the vacuole simply as the cell's "trash can". However, recent evidence suggest that the vacuole is a highly regulated structure, which carefully maintains stores of specific free sugars and amino acids, and is critical to cellular pH homeostasis, mitochondrial function, and replicative life span in yeast [109, 110]. Much of the specificity in this process results from the numerous transporters localized to the vacuole, which selectively transport individual sugars, amino acids, ions, and other species. Intriguingly, although many have been identified and cloned, some are simply hypothetical based on electrophysiology studies [111]. A better understanding of this metabolic potential will be essential for future metabolic engineering efforts.

One metabolite whose vacuolar accumulation has been exploited for metabolic engineering purposes is *S*-adenosyl methionine (SAM). SAM is the principal cellular currency for methyl transfer reactions and is a key cofactor in numerous enzymatic reactions. The majority of cellular SAM is stored in the vacuole [112]. Recently, Bayer and colleagues undertook a metagenomic approach to identifying enzymes involved in the biosynthesis of methyl halides, industrially relevant commodity chemicals that can be upconverted to numerous other chemicals using zeolite catalysts [113]. During the initial work in *E. coli*, it was postulated that SAM concentrations were limiting production. Switching to yeast, Bayer *et al.* used a well-known targeting sequence,

the N-terminus of carboxypeptidase Y, to deliver the *Batis maritima* methyl halide transferase to the vacuole and increase productivities for methyl iodide 10-fold [114]. Further, taking advantage of the fact that SAM levels can be increased by altering media conditions, methyl iodide production was increased an additional fivefold by stimulating SAM production [115].

14.5 *De novo* Organelle Construction and Future Directions

In other cases, it may be advantageous to start with cellular structures that can be repurposed to a larger extent and, perhaps, to create organelle function *de novo*. The simplest version of this idea is to completely hijack an existing organelle with less essential function. For example, peroxisomes are oxidative organelles that sequester the toxic reactions methyltrophy and/or very-long-chain fatty acid catabolism, but are not required for cellular viability under most conditions [116]. As such, they are an intriguing target for engineering. Understanding peroxisome biogenesis is still a work in progress, but proteomics experiments indicate the peroxisome contains about 10× fewer proteins than the mitochondrion, lending credence to the idea of simplicity [117]. Importantly, there are also well-defined targeting signals to both the matrix of the peroxisome and the membrane [118, 119]. The biosynthetic capability of the peroxisome has been exploited to improved production of biofuels such as fatty alcohols [120, 121] and alkanes [121].

A more ambitious area of research is to construct an entire organelle-like structure *de novo*. From a materials science perspective, organelles are formed when a set of molecular building blocks spontaneously self-organize, through molecular interactions, into complex patterns [24]. *De novo* design therefore requires identifying and engineering self-organizing building blocks. Additional properties that emerge from this process are organelle size/shape and copy number. These too must be accounted for. Recent work from Lim and colleagues demonstrates how this may be possible [122]. By leveraging the various lipid binding and lipid synthesis/degradation domains from the phosphatidylinositol signaling pathway, coupled with positive and negative feedback, Chau *et al.* were able to create pole-localized lipid microdomains. Given the large toolkit of phosphatidylinositol-binding domains, these microdomains could serve as the initial scaffold for generating more complex structures. In an orthogonal approach, Eriksson and colleagues demonstrated that overexpression of an integral membrane lipid glycosyltransferase yields massive vesicle formation in *E. coli* [123]. Combining these approaches may enable the creation of targetable distinct lipid-bound structures with controllable size and copy number, although this remains a tremendous challenge. However, such a synthetic organelle may also help to shed light on the natural organelle biogenesis process [24].

Finally, it may be revealing to reflect upon even more complex engineering challenges. Biology clearly takes advantage of compartmentalization far beyond a single genome. For example, the sea slug *Elysia chlorotica* carries out

kleptoplasty – the theft of an organelle – and spends much of its life living not as an animal but as a plant after acquiring its algal prey, *Vaucheria litorea* [124]. This intra-corpus symbiosis is maintained for the life of the slug by a yet unexplained mechanism [124]. It has been proposed this mechanism may involve the exchange of genetic material, but there are conflicting reports from differing experimental modalities as to whether algal DNA is actually incorporated into *Elysia's* genome [125, 126]. This inspires a remarkable research question: can kleptoplasty and endosymbiosis be engineered? It remains to be seen, but work from Silver and colleagues is an intriguing first step. Agapakis *et al.* found that cyanobacteria are surprisingly innocuous and do little to disturb viability when injected into zebrafish embryos [127]. Even more surprising, cyanobacteria expressing invasin and listeriolysin can grow and divide, intracellularly, in macrophages while generating little to no immunogenic response.

Alternatively, one could also imagine engineering extracellular symbioses. For example, one obvious use would be in biofuel production. There is considerable interest in constructing an organism for consolidated bioprocessing of plant-based biomass into fuel. This would entail engineering an organism for both fuel production and cellulose degradation, the major component of plant-based biomass [128]. An alternative to this approach would be to develop a stable coculture of two or more organisms that accomplish the same thing. This would potentially be more modular as the chemical production pathway remains independent of sugar consumption. Interestingly, it has been found that stable communities composed of just a handful of bacterial species can indeed degrade cellulose [129]. Moreover, recent work using *E. coli* also suggests that stable mutualism can be predicted using metabolic flux modeling, which could help systematize future engineering efforts [130]. As a demonstration of this bottom-up strategy, Mee *et al.* designed and constructed a 14-member consortium of *E. coli* mutants that were able to survive up to 50 days, although it was ultimately dominated by only four strains [131]. In addition to single-species cocultures, it is possible to engineer stable mutualism between multiple species of microbes. For example, a synthetic fungal–bacterial consortium consisting of lignocellulose-degrading *Trichoderma reesei* and an isobutanol-producing *E. coli* strain can produce the branched alcohol from corn stover [132]. Most recently, Hays and colleagues paired a heterotroph such as *E. coli*, *Bacillus subtilis*, or *S. cerevisiae* with the cyanobacteria *Synechococcus elongatus* PCC7942 mutant that has increased sucrose exporting ability. The resulting synthetic consortia can survive for a long period of time (weeks to months). In addition, by changing the heterotroph, production of large quantity of enzyme amylase (in the case of *B. subtilis*) and polyhydroxybutanoate (PHB) (*E. coli*) could be achieved [133]. Therefore, this symbiotic platform shows promise as a new modular strategy for capturing light energy in the form of bioproducts. It remains to be seen, however, how tractable these communities will be to engineering to what extent they can be scaled up in an industrial setting. Nevertheless, this represents an interesting and untapped avenue for synthetic biology, perhaps informing both our understanding of endosymbiosis and the evolution of the cell as well as the interspecies interactions central to ecology.

Acknowledgments

We thank Poh Kheng Teng for critical reading of the manuscript. This work was supported by the Department of Energy, Office of Science Early Career Research Program, through Office of Basic Energy Sciences Grant DE-SC0006394 with additional support from the Energy Biosciences Institute and the Alfred P. Sloan Foundation.

References

1 Mitchell, P. (1961) Coupling of phosphorylation to electron and hydrogen transfer by a chemi-osmotic type of mechanism. *Nature*, **191**, 144–148.
2 Voet, D. and Voet, J.G. (2011) *Biochemistry*, 4th edn, John Wiley & Sons, Inc., Hoboken, NJ.
3 Opperdoes, F.R. (2010) in *The Glycosome of Trypanosomatids* (ed. W. de Souza), Springer-Verlag, Berlin, Heidelberg, pp. 285–298. Retrieved from: http://link.springer.com/chapter/10.1007/978-3-642-12863-9_12/fulltext.html.
4 Haanstra, J.R., van Tuijl, A., Kessler, P., Reijnders, W., Michels, P.A.M., Westerhoff, H.V., Parsons, M., and Bakker, B.M. (2008) Compartmentation prevents a lethal turbo-explosion of glycolysis in trypanosomes. *Proc. Natl. Acad. Sci. U.S.A.*, **105** (46), 17718–17723.
5 Weeks, A., Lund, L., and Raushel, F.M. (2006) Tunneling of intermediates in enzyme-catalyzed reactions. *Curr. Opin. Chem. Biol.*, **10** (5), 465–472.
6 Hyde, C.C., Ahmed, S.A., Padlan, E.A., Miles, E.W., and Davies, D.R. (1988) Three-dimensional structure of the tryptophan synthase $\alpha_2\beta_2$ multienzyme complex from *Salmonella typhimurium*. *J. Biol. Chem.*, **263** (33), 17857–17871.
7 Darnault, C., Volbeda, A., Kim, E.J., Legrand, P., Vernède, X., Lindahl, P.A., and Fontecilla-Camps, J.C. (2003) Ni-Zn-[Fe_4-S_4] and Ni-Ni-[Fe_4-S_4] clusters in closed and open subunits of acetyl-CoA synthase/carbon monoxide dehydrogenase. *Nat. Struct. Biol.*, **10** (4), 271–279.
8 Thoden, J.B., Holden, H.M., Wesenberg, G., Raushel, F.M., and Rayment, I. (1997) Structure of carbamoyl phosphate synthetase: a journey of 96 A from substrate to product. *Biochemistry*, **36** (21), 6305–6316.
9 Yeates, T.O., Crowley, C.S., and Tanaka, S. (2010) Bacterial microcompartment organelles: protein shell structure and evolution. *Annu. Rev. Biophys.*, **39** (1), 185–205.
10 Penrod, J.T. and Roth, J.R. (2006) Conserving a volatile metabolite: a role for carboxysome-like organelles in *Salmonella enterica*. *J. Bacteriol.*, **188** (8), 2865–2874.
11 Sampson, E.M. and Bobik, T.A. (2008) Microcompartments for B_{12}-dependent 1,2-propanediol degradation provide protection from DNA and cellular damage by a reactive metabolic intermediate. *J. Bacteriol.*, **190** (8), 2966–2971.
12 O'Connell, J.D., Zhao, A., Ellington, A.D., and Marcotte, E.M. (2012) Dynamic reorganization of metabolic enzymes into intracellular bodies. *Annu. Rev. Cell Dev. Biol.*, **28**, 89–111.

13 Keasling, J.D. (2008) Synthetic biology for synthetic chemistry. *ACS Chem. Biol.*, **3** (1), 64–76.

14 Savage, D.F., Way, J.C., and Silver, P.A. (2008) Defossiling fuel: how synthetic biology can transform biofuel production. *ACS Chem. Biol.*, **3** (1), 13–16.

15 Dueber, J.E., Wu, G.C., Malmirchegini, G.R., Moon, T.S., Petzold, C.J., Ullal, A.V., Prather, K.L.J., and Keasling, J.D. (2009) Synthetic protein scaffolds provide modular control over metabolic flux. *Nat. Biotechnol.*, **27** (8), 753–759.

16 Sachdeva, G., Garg, A., Godding, D., Way, J.C., and Silver, P.A. (2014) In vivo co-localization of enzymes on RNA scaffolds increases metabolic production in a geometrically dependent manner. *Nucleic Acids Res.*, **42**, 9493–9503.

17 Agapakis, C.M., Boyle, P.M., and Silver, P.A. (2012) Natural strategies for the spatial optimization of metabolism in synthetic biology. *Nat. Chem. Biol.*, **8** (6), 527–535.

18 Lee, H., DeLoache, W.C., and Dueber, J.E. (2012) Spatial organization of enzymes for metabolic engineering. *Metab. Eng.*, **14** (3), 242–251.

19 Chen, A.H. and Silver, P.A. (2012) Designing biological compartmentalization. *Trends Cell Biol.*, **22** (12), 662–670.

20 Diekmann, Y. and Pereira-Leal, J.B. (2013) Evolution of intracellular compartmentalization. *Biochem. J*, **449** (2), 319–331.

21 Mullock, B.M. and Luzio, J.P. (2005) Theory of organelle biogenesis, in *The Biogenesis of Cellular Organelles*, Springer, Boston, MA, pp. 1–18.

22 Liberton, M., Berg, R.H., Heuser, J., Roth, R., and Pakrasi, H.B. (2006) Ultrastructure of the membrane systems in the unicellular cyanobacterium *Synechocystis* sp. strain PCC 6803. *Protoplasma*, **227** (2–4), 129–138.

23 Mangan, N.M., Flamholz, A., Hood, R.D., Milo, R., and Savage, D.F. (2016) pH determines the energetic efficiency of the cyanobacterial CO_2 concentrating mechanism. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 5354–5362. doi: 10.1073/pnas.1525145113

24 Rafelski, S.M. and Marshall, W.F. (2008) Building the cell: design principles of cellular architecture. *Nat. Rev. Mol. Cell Biol.*, **9** (8), 593–602.

25 Gaertner, F.H. (1978) Unique catalytic properties of enzyme clusters. *Trends Biochem. Sci*, **3** (1), 63–65.

26 Conrado, R.J., Varner, J.D., and DeLisa, M.P. (2008) Engineering the spatial organization of metabolic enzymes: mimicking nature's synergy. *Curr. Opin. Biotechnol.*, **19** (5), 492–499.

27 Kerfeld, C.A., Heinhorst, S., and Cannon, G.C. (2010) Bacterial microcompartments. *Annu. Rev. Microbiol.*, **64** (1), 391–408.

28 Cannon, G.C., Bradburne, C.E., Aldrich, H.C., Baker, S.H., Heinhorst, S., and Shively, J.M. (2001) Microcompartments in prokaryotes: carboxysomes and related polyhedra. *Appl. Environ. Microbiol.*, **67** (12), 5351–5361.

29 Axen, S.D., Erbilgin, O., and Kerfeld, C.A. (2014) A taxonomy of bacterial microcompartment loci constructed by a novel scoring method. *PLoS Comput. Biol.*, **10**, e1003898.

30 Dou, Z., Heinhorst, S., Williams, E.B., Murin, C.D., Shively, J.M., and Cannon, G.C. (2008) CO_2 fixation kinetics of *Halothiobacillus neapolitanus* mutant carboxysomes lacking carbonic anhydrase suggest the shell acts as a diffusional barrier for CO_2. *J. Biol. Chem.*, **283** (16), 10377–10384.

31 Shih, P.M., Zarzycki, J., Niyogi, K.K., and Kerfeld, C.A. (2014) Introduction of a synthetic CO_2-fixing photorespiratory bypass into a cyanobacterium. *J. Biol. Chem.*, **289** (14), 9493–9500.

32 Espie, G.S. and Kimber, M.S. (2011) Carboxysomes: cyanobacterial RubisCO comes in small packages. *Photosynth. Res.*, **109**, 7–20.

33 Chowdhury, C., Sinha, S., Chun, S., Yeates, T.O., and Bobik, T.A. (2014) Diverse bacterial microcompartment organelles. *Microbiol. Mol. Biol. Rev.*, **78**, 438–468.

34 Kerfeld, C.A., Sawaya, M.R., Tanaka, S., Nguyen, C.V., Phillips, M., Beeby, M., and Yeates, T.O. (2005) Protein structures forming the shell of primitive bacterial organelles. *Science*, **309** (5736), 936–938.

35 Yeates, T.O., Thompson, M.C., and Bobik, T.A. (2011) The protein shells of bacterial microcompartment organelles. *Curr. Opin. Struct. Biol.*, **21** (2), 223–231.

36 Tanaka, S., Kerfeld, C.A., Sawaya, M.R., Cai, F., Heinhorst, S., Cannon, G.C., and Yeates, T.O. (2008) Atomic-level models of the bacterial carboxysome shell. *Science*, **319** (5866), 1083–1086.

37 Iancu, C.V., Ding, H.J., Morris, D.M., Dias, D.P., Gonzales, A.D., Martino, A., and Jensen, G.J. (2007) The structure of isolated *Synechococcus* strain WH8102 carboxysomes as revealed by electron cryotomography. *J. Mol. Biol.*, **372** (3), 764–773.

38 Schmid, M.F., Paredes, A.M., Khant, H.A., Soyer, F., Aldrich, H.C., Chiu, W., and Shively, J.M. (2006) Structure of *Halothiobacillus neapolitanus* carboxysomes by cryo-electron tomography. *J. Mol. Biol.*, **364** (3), 526–535.

39 Wheatley, N.M., Gidaniyan, S.D., Liu, Y., Cascio, D., and Yeates, T.O. (2013) Bacterial microcompartment shells of diverse functional types possess pentameric vertex proteins. *Protein Sci.*, **22** (5), 660–665.

40 Shively, J.M., Ball, F.L., and Kline, B.W. (1973) Electron microscopy of the carboxysomes (polyhedral bodies) of *Thiobacillus neapolitanus*. *J. Bacteriol.*, **116** (3), 1405–1411.

41 Kaplan, A. and Reinhold, L. (1999) CO_2 concentrating mechanisms in photosynthetic microorganisms. *Annu. Rev. Plant Physiol. Plant Mol. Biol.*, **50**, 539–570.

42 Badger, M.R. and Price, G.D. (2003) CO_2 concentrating mechanisms in cyanobacteria: molecular components, their diversity and evolution. *J. Exp. Bot. Soc. Exp. Biol*, **54** (383), 609–622.

43 Castellana, M., Wilson, M.Z., Xu, Y., Joshi, P., Cristea, I.M., Rabinowitz, J.D. *et al.* (2014) Enzyme clustering accelerates processing of intermediates through metabolic channeling. *Nat. Biotechnol.*, **32**, 1011–1018.

44 Yeates, T.O., Kerfeld, C.A., Heinhorst, S., Cannon, G.C., and Shively, J.M. (2008) Protein-based organelles in bacteria: carboxysomes and related microcompartments. *Nat. Rev. Microbiol.*, **6** (9), 681–691.

45 Whitehead, L., Long, B.M., Price, G.D., and Badger, M.R. (2014) Comparing the in vivo function of α-carboxysomes and β-carboxysomes in two model cyanobacteria. *Plant Physiol.*, **165**, 398–411.

46 Fan, C., Cheng, S., Liu, Y., Escobar, C.M., Crowley, C.S., Jefferson, R.E., Yeates, T.O., and Bobik, T.A. (2010) Short N-terminal sequences package proteins into bacterial microcompartments. *Proc. Natl. Acad. Sci. U.S.A.*, **107** (16), 7509–7514.

47 Lawrence, A.D., Frank, S., Newnham, S., Lee, M.J., Brown, I.R., Xue, W.-F. *et al.* (2014) Solution structure of a bacterial microcompartment targeting peptide and its application in the construction of an ethanol bioreactor. *ACS Synth. Biol.*, **3** (7), 454–465.

48 Fan, C., Cheng, S., Sinha, S., and Bobik, T.A. (2012) Interactions between the termini of lumen enzymes and shell proteins mediate enzyme encapsulation into bacterial microcompartments. *Proc. Natl. Acad. Sci. U.S.A.*, **109** (37), 14995–15000.

49 Fan, C. and Bobik, T.A. (2011) The N-terminal region of the medium subunit (PduD) packages adenosylcobalamin-dependent diol dehydratase (PduCDE) into the Pdu microcompartment. *J. Bacteriol.*, **193**, 5623–5628.

50 Parsons, J.B., Frank, S., Bhella, D., Liang, M., Prentice, M.B., Mulvihill, D.P. *et al.* (2010) Synthesis of empty bacterial microcompartments, directed organelle protein incorporation, and evidence of filament-associated organelle movement. *Mol. Cell*, **38**, 305–315. doi: 10.1016/j.molcel.2010.04.008

51 Choudhary, S., Quin, M.B., Sanders, M.A., Johnson, E.T. *et al.* (2012) Engineered protein nano-compartments for targeted enzyme localization. *PLoS One*, **7**, e33342.

52 Quin, M.B., Perdue, S.A., Hsu, S., and Schmidt-Dannert, C. (2016) Encapsulation of multiple cargo proteins within recombinant Eut nanocompartments. *Appl. Microbiol. Biotechnol.*, **100**, 9187–9200.

53 Jakobson, C.M., Slininger Lee, M.F., and Tullman-Ercek, D. (2017) *De novo* design of signal sequences to localize cargo to the 1,2-propanediol utilization microcompartment. *Protein Sci.*, **22**, 1–50. doi: 10.1002/pro.3144

54 Liang, M., Frank, S., Lünsdorf, H., Warren, M.J., and Prentice, M.B. (2017) Bacterial microcompartment-directed polyphosphate kinase promotes stable polyphosphate accumulation in *E. coli*. *Biotechnol. J.*, 1600415. doi: 10.1002/biot.201600415

55 Cameron, J.C., Wilson, S.C., Bernstein, S.L., and Kerfeld, C.A. (2013) Biogenesis of a bacterial organelle: the carboxysome assembly pathway. *Cell*, **155**, 1131–1140.

56 Kinney, J.N., Salmeen, A., Cai, F., and Kerfeld, C.A. (2012) Elucidating essential role of conserved carboxysomal protein CcmN reveals common feature of bacterial microcompartment assembly. *J. Biol. Chem.*, **287** (21), 17729–17736.

57 Cai, F., Dou, Z., Bernstein, S., Leverenz, R., Williams, E., Heinhorst, S. *et al.* (2015) Advances in understanding carboxysome assembly in *Prochlorococcus* and *Synechococcus* implicate CsoS2 as a critical component. *Life*, **5**, 1141–1171.

58 Chaijarasphong, T., Nichols, R.J., Kortright, K.E., Nixon, C.F., Teng, P.K., Oltrogge, L.M. *et al.* (2016) Programmed ribosomal frameshifting mediates expression of the α-carboxysome. *J. Mol. Biol.*, **428**, 153–164.

59 Tompa, P. (2012) Intrinsically disordered proteins: a 10-year recap. *Trends Biochem. Sci*, **37**, 509–516.

60 Menon, B.B., Dou, Z., Heinhorst, S., Shively, J.M., and Cannon, G.C. (2008) Halothiobacillus neapolitanus carboxysomes sequester heterologous and chimeric RubisCO species. (ed. J. Rutherford) . *PLoS One*, **3** (10), e3570.

61 Bonacci, W., Teng, P.K., Afonso, B., Niederholtmeyer, H., Grob, P., Silver, P.A., and Savage, D.F. (2012) Modularity of a carbon-fixing protein organelle. *Proc. Natl. Acad. Sci. U.S.A.*, **109** (2), 478–483.

62 Kinney, J.N., Axen, S.D., and Kerfeld, C.A. (2011) Comparative analysis of carboxysome shell proteins. *Photosynth. Res.*, **109**, 21–32. doi: 10.1007/s11120-011-9624-6

63 Young, J.D., Shastri, A.A., Stephanopoulos, G., and Morgan, J.A. (2011) Mapping photoautotrophic metabolism with isotopically nonstationary (13)C flux analysis. *Metab. Eng.*, **13** (6), 656–665.

64 Marcus, Y., Berry, J.A., and Pierce, J. (1992) Photosynthesis and photorespiration in a mutant of the cyanobacterium *Synechocystis* PCC 6803 lacking carboxysomes. *Planta*, **187** (4), 511–516.

65 Cai, F., Menon, B.B., Cannon, G.C., Curry, K.J., Shively, J.M., and Heinhorst, S. (2009) The pentameric vertex proteins are necessary for the icosahedral carboxysome shell to function as a CO_2 leakage barrier. (ed. N. Ahmed) . *PLoS One*, **4** (10), e7521.

66 Chowdhury, C., Chun, S., Pang, A., Sawaya, M.R., Sinha, S., Yeates, T.O., and Bobik, T.A. (2015) Selective molecular transport through the protein shell of a bacterial microcompartment organelle. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 2990–2995.

67 Klein, M.G., Zwart, P., Bagby, S.C., Cai, F., Chisholm, S.W., Heinhorst, S., Cannon, G.C., and Kerfeld, C.A. (2009) Identification and structural analysis of a novel carboxysome shell protein with implications for metabolite transport. *J. Mol. Biol.*, **392** (2), 319–333.

68 Roberts, E.W., Cai, F., Kerfeld, C.A., Cannon, G.C., and Heinhorst, S. (2012) Isolation and characterization of the *Prochlorococcus* carboxysome reveal the presence of the novel shell protein CsoS1D. *J. Bacteriol.*, **194** (4), 787–795.

69 Cai, F., Sutter, M., Cameron, J.C., Stanley, D.N., Kinney, J.N., and Kerfeld, C.A. (2013) The structure of CcmP, a tandem bacterial microcompartment domain protein from the β-carboxysome, forms a subcompartment within a microcompartment. *J. Biol. Chem.*, **288**, 16055–16063.

70 Thompson, M.C., Cascio, D., Leibly, D.J., and Yeates, T.O. (2015) An allosteric model for control of pore opening by substrate binding in the EutL microcompartment shell protein. *Protein Sci.*, **24**, 956–975. doi: 10.1002/pro.2672

71 Parsons, J.B., Dinesh, S.D., Deery, E., Leech, H.K., Brindley, A.A., Heldt, D., Frank, S., Smales, C.M., Lünsdorf, H., Rambach, A., Gass, M.H., Bleloch, A., McClean, K.J., Munro, A.W., Rigby, S.E.J., Warren, M.J., and Prentice, M.B. (2008) Biochemical and structural insights into bacterial organelle form and biogenesis. *J. Biol. Chem.*, **283** (21), 14366–14375.

72 Crowley, C.S., Cascio, D., Sawaya, M.R., Kopstein, J.S., Bobik, T.A., and Yeates, T.O. (2010) Structural insight into the mechanisms of transport across the *Salmonella enterica* Pdu microcompartment shell. *J. Biol. Chem.*, **285** (48), 37838–37846.

73 Thompson, M.C., Wheatley, N.M., Jorda, J., Sawaya, M.R., Gidaniyan, S.D., Ahmed, H. *et al.* (2014) Identification of a unique Fe-S cluster binding site in a glycyl-radical type microcompartment shell protein. *J. Mol. Biol.*, **426**, 3287–3304. doi: 10.1016/j.jmb.2014.07.018

74 Chen, A.H., Robinson-Mosher, A., Savage, D.F., Silver, P.A., and Polka, J.K. (2013) The bacterial carbon-fixing organelle is formed by shell envelopment of preassembled cargo. *PLoS One*, **8**, e76127.

75 Heinhorst, S., Cannon, G.C., and Shively, J.M. (2006) Carboxysomes and carboxysome-like inclusions, in *Complex Intracellular Structures in Prokaryotes*, Microbiology Monographs (ed. J.M. Shively), Springer-Verlag, Berlin, pp. 141–164.

76 Milo, R., Jorgensen, P., Moran, U., Weber, G., and Springer, M. (2010) BioNumbers – the database of key numbers in molecular and cell biology. *Nucleic Acids Res.*, **38** (Database issue), D750–D753.

77 Chowdhury, C., Chun, S., Sawaya, M.R., Yeates, T.O., and Bobik, T.A. (2016) The function of the PduJ microcompartment shell protein is determined by the genomic position of its encoding gene. *Mol. Microbiol.*, **101**, 770–783. doi: 10.1111/mmi.13423

78 Havemann, G.D. and Bobik, T.A. (2003) Protein content of polyhedral organelles involved in coenzyme B_{12}-dependent degradation of 1,2-propanediol in *Salmonella enterica* serovar Typhimurium LT2. *J. Bacteriol.*, **185** (17), 5086–5095.

79 Lin, M.T., Occhialini, A., Andralojc, P.J., Devonshire, J., Hines, K.M., Parry, M.A.J. *et al.* (2014) β-Carboxysomal proteins assemble into highly organized structures in Nicotiana chloroplasts. *Plant J.*, **79**, 1–12.

80 Douglas, T. and Young, M. (2006) Viruses: making friends with old foes. *Science*, **312** (5775), 873–875.

81 Rome, L.H. and Kickhoefer, V.A. (2013) Development of the vault particle as a platform technology. *ACS Nano*, **7** (2), 889–902.

82 Worsdorfer, B., Woycechowsky, K.J., and Hilvert, D. (2011) Directed evolution of a protein container. *Science*, **331** (6017), 589–592.

83 Sutter, M., Boehringer, D., Gutmann, S., Günther, S., Prangishvili, D., Loessner, M.J., Stetter, K.O., Weber-Ban, E., and Ban, N. (2008) Structural basis of enzyme encapsulation into a bacterial nanocompartment. *Nat. Struct. Mol. Biol.*, **15** (9), 939–947.

84 Bode, S.A., Minten, I.J., Nolte, R.J.M., and Cornelissen, J.J.L.M. (2011) Reactions inside nanoscale protein cages. *Nanoscale*, **3** (6), 2376–2389.

85 Maity, B., Fujita, K., and Ueno, T. (2015) Use of the confined spaces of apoferritin and virus capsids as nanoreactors for catalytic reactions. *Curr. Opin. Chem. Biol.*, **25**, 88–97. doi: 10.1016/j.cbpa.2014.12.026

86 Valdés-Stauber, N. and Scherer, S. (1994) Isolation and characterization of Linocin M18, a bacteriocin produced by *Brevibacterium linens*. *Appl. Environ. Microbiol. Am. Soc. Microbiol.*, **60** (10), 3809–3814.

87 Giessen, T.W. and Silver, P.A. (2017) Widespread distribution of encapsulin nanocompartments reveals functional diversity. *Nat. Microbiol.*, **2**, 17029. doi: 10.1038/nmicrobiol.2017.29

88 Radford, D.R. (2014) Understanding the encapsulins: prediction and characterization of phage capsid-like nanocompartments in prokaryotes. University of Toronto. Dissertation.

89 Akita, F., Chong, K.T., Tanaka, H., Yamashita, E., Miyazaki, N., Nakaishi, Y., Suzuki, M., Namba, K., Ono, Y., Tsukihara, T., and Nakagawa, A. (2007) The crystal structure of a virus-like particle from the hyperthermophilic archaeon *Pyrococcus furiosus* provides insight into the evolution of viruses. *J. Mol. Biol.*, **368** (5), 1469–1483.

90 Tamura, A. *et al.* (2015) Packaging guest proteins into the encapsulin nanocompartment from *Rhodococcus erythropolis* N771. *Biotechnol. Bioeng.*, **112** (1), 13–20.

91 Cassidy-Amstutz, C. (2016) Identification of a minimal peptide tag for in vivo and in vitro loading of encapsulin. *Biochemistry*, **55** (24), 3461–3468.

92 Snijder, J. *et al.* (2015) Defining the stoichiometry and cargo load of viral and bacterial nanoparticles by Orbitrap mass spectrometry. *J. Am. Chem. Soc.*, **136** (20), 7295–7299.

93 Rahmanpour, R. and Bugg, T.D.H. (2013) Assembly in vitro of *Rhodococcus jostii* RHA1 encapsulin and peroxidase DypB to form a nanocompartment. *FEBS J.*, **280** (9), 2097–2104.

94 Snijder, J. *et al.* (2016) Assembly and mechanical properties of the cargo-free and cargo-loaded bacterial nanocompartment encapsulin. *Biomacromolecules*, **17** (8), 2522–2529.

95 Calvo, S.E. and Mootha, V.K. (2010) The mitochondrial proteome and human disease. *Annu. Rev. Genomics Hum. Genet.*, **11**, 25–44.

96 Llopis, J., McCaffery, J.M., Miyawaki, A., Farquhar, M.G., and Tsien, R.Y. (1998) Measurement of cytosolic, mitochondrial, and Golgi pH in single living cells with green fluorescent proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **95** (12), 6803–6808.

97 Hiltunen, J.K., Schonauer, M.S., Autio, K.J., Mittelmeier, T.M., Kastaniotis, A.J., and Dieckmann, C.L. (2009) Mitochondrial fatty acid synthesis type II: more than just fatty acids. *J. Biol. Chem.*, **284** (14), 9011–9015.

98 Pagliarini, D.J., Calvo, S.E., Chang, B., Sheth, S.A., Vafai, S.B., Ong, S.-E., Walford, G.A., Sugiana, C., Boneh, A., Chen, W.K., Hill, D.E., Vidal, M., Evans, J.G., Thorburn, D.R., Carr, S.A., and Mootha, V.K. (2008) A mitochondrial protein compendium elucidates complex I disease biology. *Cell*, **134** (1), 112–123.

99 Keasling, J.D. (2010) Microbial production of isoprenoids, in *Handbook of Hydrocarbon and Lipid Microbiology*, Springer-Verlag, Berlin Heidelberg.

100 Farhi, M., Marhevka, E., Masci, T., Marcos, E., Eyal, Y., Ovadis, M., Abeliovich, H., and Vainstein, A. (2011) Harnessing yeast subcellular compartments for the production of plant terpenoids. *Metab. Eng.*, **13** (5), 474–481.

101 Hurt, E.C., Müller, U., and Schatz, G. (1985) The first twelve amino acids of a yeast mitochondrial outer membrane protein can direct a nuclear-coded cytochrome oxidase subunit to the mitochondrial inner membrane. *EMBO J.*, **4** (13A), 3509.

102 Phillips, M.A., D'Auria, J.C., Gershenzon, J., and Pichersky, E. (2008) The Arabidopsis thaliana type I isopentenyl diphosphate isomerases are targeted to multiple subcellular compartments and have overlapping functions in isoprenoid biosynthesis. *Plant Cell*, **20** (3), 677–696.

103 Hazelwood, L.A., Daran, J.-M., van Maris, A.J.A., Pronk, J.T., and Dickinson, J.R. (2008) The Ehrlich pathway for fusel alcohol production: a century of research on *Saccharomyces cerevisiae* metabolism. *Appl. Environ. Microbiol.*, **74** (8), 2259–2266.

104 Avalos, J.L., Fink, G.R., and Stephanopoulos, G. (2013) Compartmentalization of metabolic pathways in yeast mitochondria improves the production of branched-chain alcohols. *Nat. Biotechnol.*, **31** (4), 335–341.

105 Li, S., Liu, L., and Chen, J. (2015) Compartmentalizing metabolic pathway in *Candida glabrata* for acetoin production. *Metab. Eng.*, **28**, 1–7. doi: 10.1016/j.ymben.2014.11.008

106 Chen, X., Zhu, P., and Liu, L. (2016) Modular optimization of multi-gene pathways for fumarate production. *Metab. Eng.*, **33**, 76–85.

107 Malhotra, K., Subramaniyan, M., Rawat, K., Kalamuddin, M., Qureshi, M.I., Malhotra, P. *et al.* (2016) Compartmentalized metabolic engineering for artemisinin biosynthesis and effective malaria treatment by oral delivery of plant cells. *Mol. Plant*, **9**, 1464–1477. doi: 10.1016/j.molp.2016.09.013

108 Klionsky, D.J., Herman, P.K., and Emr, S.D. (1990) The fungal vacuole: composition, function, and biogenesis. *Microbiol. Rev.*, **54** (3), 266.

109 Oikawa, A., Matsuda, F., Kikuyama, M., Mimura, T., and Saito, K. (2011) Metabolomics of a single vacuole reveals metabolic dynamism in an alga *Chara australis*. *Plant Physiol.*, **157** (2), 544–551.

110 Hughes, A.L. and Gottschling, D.E. (2013) An early age increase in vacuolar pH limits mitochondrial function and lifespan in yeast. *Nature*, **492** (7428), 261–265.

111 Martinoia, E., Maeshima, M., and Neuhaus, H.E. (2007) Vacuolar transporters and their essential role in plant metabolism. *J. Exp. Bot.*, **58** (1), 83–102.

112 Farooqui, J.Z., Lee, H.W., Kim, S., and Paik, W.K. (1983) Studies on compartmentation of *S*-adenosyl-l-methionine in *Saccharomyces cerevisiae* and isolated rat hepatocytes. *Biochim. Biophys. Acta*, **757** (3), 342–351.

113 Bayer, T.S., Widmaier, D.M., Temme, K., Mirsky, E.A., Santi, D.V., and Voigt, C.A. (2009) Synthesis of methyl halides from biomass using engineered microbes. *J. Am. Chem. Soc.*, **131** (18), 6508–6515.

114 Valls, L.A., Winther, J.R., and Stevens, T.H. (1990) Yeast carboxypeptidase Y vacuolar targeting signal is defined by four propeptide amino acids. *J. Cell Biol.*, **111** (2), 361–368.

115 Lin, J.-P., Tian, J., You, J.-F., Jin, Z.-H., Xu, Z.-N., and Cen, P.-L. (2004) An effective strategy for the co-production of *S*-adenosyl-l-methionine and glutathione by fed-batch fermentation. *Biochem. Eng. J.*, **21** (1), 19–25.

116 van der Klei, I.J. and Veenhuis, M. (2006) Yeast and filamentous fungi as model organisms in microbody research. *Biochim. Biophys. Acta*, **1763** (12), 1364–1373.

117 Saleem, R.A., Smith, J.J., and Aitchison, J.D. (2006) Proteomics of the peroxisome. *Biochim. Biophys. Acta*, **1763** (12), 1541–1551.

118 Léon, S., Goodman, J.M., and Subramani, S. (2006) Uniqueness of the mechanism of protein import into the peroxisome matrix: transport of folded, co-factor-bound and oligomeric proteins by shuttling receptors. *Biochim. Biophys. Acta*, **1763** (12), 1552–1564.

119 DeLoache, W.C., Russ, Z.N., and Dueber, J.E. (2016) Towards repurposing the yeast peroxisome for compartmentalizing heterologous metabolic pathways. *Nat. Commun.*, **7**, 11152. doi: 10.1038/ncomms11152

120 Sheng, J., Stevens, J., and Feng, X. (2016) Pathway compartmentalization in peroxisome of *Saccharomyces cerevisiae* to produce versatile medium chain fatty alcohols. *Sci. Rep.*, **6**, 26884. doi: 10.1038/srep26884

121 Zhou, Y.J., Buijs, N.A., Zhu, Z., Gomez, D.O., Boonsombuti, A., Siewers, V. *et al.* (2016) Harnessing yeast peroxisomes for biosynthesis of fatty-acid-derived biofuels and chemicals with relieved side-pathway competition. *J. Am. Chem. Soc.*, **138**, 15368–15377. doi: 10.1021/jacs.6b07394

122 Chau, A.H., Walter, J.M., Gerardin, J., Tang, C., and Lim, W.A. (2012) Designing synthetic regulatory networks capable of self-organizing cell polarization. *Cell*, **151** (2), 320–332.

123 Eriksson, H.M., Wessman, P., Ge, C., Edwards, K., and Wieslander, Å. (2009) Massive formation of intracellular membrane vesicles in *Escherichia coli* by a monotopic membrane-bound lipid glycosyltransferase. *J. Biol. Chem.*, **284** (49), 33904–33914.

124 Rumpho, M.E., Pelletreau, K.N., Moustafa, A., and Bhattacharya, D. (2011) The making of a photosynthetic animal. *J. Exp. Biol.*, **214** (Pt. 2), 303–311.

125 Bhattacharya, D., Pelletreau, K.N., and Price, D.C. (2013) Genome analysis of *Elysia chlorotica* egg DNA provides no evidence for horizontal gene transfer into the germ line of this kleptoplastic mollusc. *Mol. Biol. Evol.*, **30**, 1843–1852.

126 Schwartz, J.A., Curtis, N.E., and Pierce, S.K. (2014) FISH labeling reveals a horizontally transferred algal (*vaucheria litorea*) nuclear gene on a sea slug (*elysia chlorotica*) chromosome. *Biol. Bull.*, **227**, 300–312. doi: 10.1086/BBLv227n3p300

127 Agapakis, C.M., Niederholtmeyer, H., Noche, R.R., Lieberman, T.D., Megason, S.G., Way, J.C., and Silver, P.A. (2011) Towards a synthetic chloroplast. (ed. D.M. Ojcius) . *PLoS One*, **6** (4), e18877.

128 Steen, E.J., Kang, Y., Bokinsky, G., Hu, Z., Schirmer, A., McClure, A., del Cardayre, S.B., and Keasling, J.D. (2010) Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature*, **463** (7280), 559–562.

129 Kato, S., Haruta, S., Cui, Z.J., Ishii, M., and Igarashi, Y. (2005) Stable coexistence of five bacterial strains as a cellulose-degrading community. *Appl. Environ. Microbiol.*, **71** (11), 7099–7106.

130 Wintermute, E.H. and Silver, P.A. (2010) Emergent cooperation in microbial metabolism. *Mol. Syst. Biol.*, **6**, 1–7.

131 Mee, M.T., Collins, J.J., Church, G.M., and Wang, H.H. (2014) Syntrophic exchange in synthetic microbial communities. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E2149–E2156. doi: 10.1073/pnas.1405641111

132 Minty, J.J., Singer, M.E., Scholz, S.A., Bae, C.-H., Ahn, J.-H., Foster, C.E. *et al.* (2013) Design and characterization of synthetic fungal-bacterial consortia for direct production of isobutanol from cellulosic biomass. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 14592–14597. doi: 10.1073/pnas.1218447110

133 Hays, S.G., Yan, L.L.W., Silver, P.A., and Ducat, D.C. (2016) Synthetic photosynthetic consortia define interactions leading to robustness and photoproduction. *bioRxiv*, 1–23. doi: 10.1101/068130

134 Bobik, T.A. (2006) Polyhedral organelles compartmenting bacterial metabolic processes. *Appl. Microbiol. Biotechnol.*, **70** (5), 517–525.

135 Bobik, T.A., Havemann, G.D., Busch, R.J., Williams, D.S., and Aldrich, H.C. (1999) The propanediol utilization (*pdu*) operon of *salmonella enterica* serovar typhimurium LT2 includes genes necessary for formation of polyhedral

organelles involved in coenzyme B_{12}-dependent 1,2-propanediol degradation. *J. Bacteriol.*, **181** (19), 5967–5975.

136 Chen, P., Andersson, D.I., and Roth, J.R. (1994) The control region of the pdu/cob regulon in *Salmonella typhimurium*. *J. Bacteriol.*, **176** (17), 5474–5482.

137 Kofoid, E., Rappleye, C., Stojiljkovic, I., and Roth, J. (1999) The 17-gene ethanolamine (eut) operon of *salmonella typhimurium* encodes five homologues of carboxysome shell proteins. *J. Bacteriol.*, **181** (17), 5317–5329.

138 Peña, K.L., Castel, S.E., de Araujo, C., Espie, G.S., and Kimber, M.S. (2010) Structural basis of the oxidative activation of the carboxysomal γ-carbonic anhydrase, CcmM. *Proc. Natl. Acad. Sci. U.S.A.*, **107** (6), 2455–2460.

Part IV

Early Applications of Synthetic Biology: Pathways, Therapies, and Cell-Free Synthesis

15

Cell-Free Protein Synthesis: An Emerging Technology for Understanding, Harnessing, and Expanding the Capabilities of Biological Systems

Jennifer A. Schoborg [1,2] *and Michael C. Jewett* [1,2,3,4,5]

[1] *Department of Chemical and Biological Engineering, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208-3120, USA*
[2] *Chemistry of Life Processes Institute, 2170 Campus Drive, Evanston, IL 60208-3120, USA*
[3] *Robert H. Lurie Comprehensive Cancer Center, Northwestern University, 676 N. St Clair St, Suite 1200, Chicago, IL 60611-3068, USA*
[4] *Simpson Querrey Institute, Northwestern University, 303 E. Superior St, Suite 11-131, Chicago, IL 60611-2875, USA*
[5] *Center for Synthetic Biology, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208-3120, USA*

15.1 Introduction

Cell-free protein synthesis (CFPS) systems have transformed our ability to understand, harness, and expand the capabilities of biological systems. In the groundbreaking experiments of Nirenberg and Matthaei in 1961, CFPS played an essential role in the discovery of the genetic code [1]. More recently, a technical renaissance has revitalized CFPS systems to help meet increasing demands for simple and efficient protein synthesis. Moving forward, this renaissance is enabling new processes never seen in nature, such as noncanonical amino acid (ncAA) incorporation and man-made genetic circuits.

The driving force behind this development has been the unprecedented freedom of design to modify and control biological systems that is unattainable with *in vivo* approaches [2–6]. The ability to "open the hood" of the cell and treat biology as a set of chemical reactions leads to many advantages for using cell-free systems, highlighted in Figure 15.1. First, the open reaction environment allows the user to directly influence the biochemical systems of interest (e.g., protein synthesis, metabolism, etc.). As a result, new components (natural and nonnatural) can be added or synthesized and can be maintained at precise concentrations, while the chemical environment is monitored and sampled. Second, since the reaction is not "living," cellular objectives, such as growth, can be bypassed. As is desirable in chemical transformations, cell-free systems separate catalyst synthesis (cell growth) from catalyst utilization (protein production), circumventing a major challenge afflicting *in vivo* engineering efforts. This is featured in Figure 15.2. Without living cells, timelines for process and product development can be faster and scale-up can be easier [4]. Although the CFPS technology offers many exciting advantages, challenges remain that provide

Synthetic Biology: Parts, Devices and Applications, First Edition. Edited by Christina Smolke.

CFPS gives an unprecedented freedom of design to modify and control biology

Open reaction environment

Control added components precisely

Monitor and sample reaction environment

Bypass cellular objectives

Separate catalyst synthesis from catalyst utilization

Direct resources toward the exclusive production of one product

Accelerate timelines from DNA to protein

Supplement and produce toxic molecules

Scale linearly from μl to 100 l (expansion factor of 10^6)

Figure 15.1 Advantages for cell-free biology. By bypassing cellular objectives and opening the reaction environment, cell-free protein synthesis allows for increased freedom of design as a result of the benefits highlighted here.

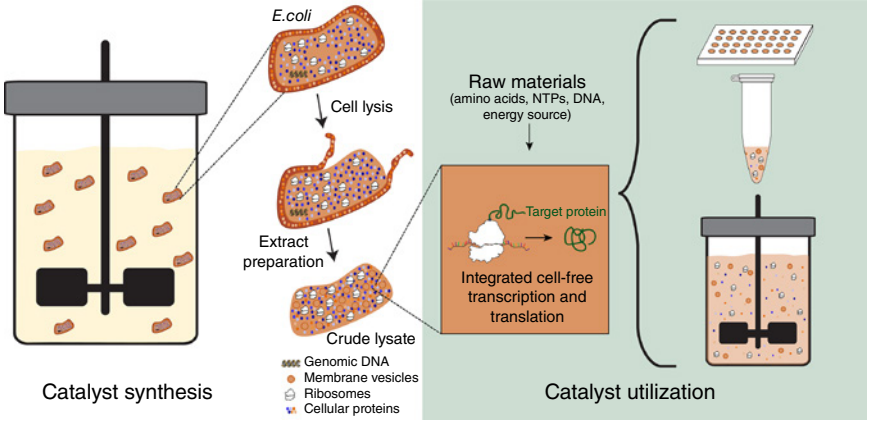


Figure 15.2 A new paradigm for cell-free biomanufacturing. Cell-free protein synthesis is able to separate catalyst synthesis (cell growth) from catalyst utilization (protein synthesis). This allows resources to be funneled toward the product of interest in ways not possible *in vivo*.

opportunity for improvement. For example, many emerging cell-free platforms are not yet commercially available, and thus their broad impact is limited. In addition, cell lysis procedures can be difficult to standardize, leading to different extract performance across labs. Further, complex posttranslational modifications (PTMs) (e.g., human glycosylation) are still limited or not yet shown. Finally, CFPS costs exceed *in vivo* methods for comparable organisms, which limit the scale for most academic labs. Despite these challenges, the benefits of CFPS are inspiring new applications from the synthesis of pharmaceutical proteins to the understanding of synthetic gene circuits [7].

This review highlights achievements of the existing systems for crude extract-based protein synthesis. We begin with an overview of the state-of-the-art systems from different organisms. Then, we discuss their capabilities for protein production, highlighting applications that greatly benefit from the open environment and lack of cell viability of CFPS. Finally, we describe benefits for high-throughput applications and offer some commentary about the future growth of the field.

15.2 Background/Current Status

Crude extract-based CFPS harnesses the cell's native translational machinery to produce proteins in a process that, instead of occurring in a live cell, becomes more like a chemical reaction. The crude extract contains the translational machinery, which consists of ribosomes, aminoacyl-tRNA synthetases, initiation factors, elongation factors, chaperones, and so on. In addition to the translational machinery, other enzymes exist in the extract: some are beneficial (e.g., those for recycling nucleotides or energy metabolism) and some are detrimental (e.g., those using CFPS substrates nonproductively). In combined transcription–translation reactions, the crude extract is added to a solution containing buffer, amino acids, nucleotides, RNA polymerase, a secondary energy source (for regenerating adenosine triphosphate (ATP)), salts, and other molecules for maintaining the environment (e.g., dithiothreitol for a reducing environment or spermidine and putrescine for mimicking the cytoplasm). Thus far, when compared with the use of purified enzyme translation systems, such as the PURE system developed by Ueda and colleagues [8], as well as New England Biolabs [9, 10], crude cell lysates offer significantly lower system catalyst costs and much greater system capabilities (e.g., cofactor regeneration, proteins produced per ribosome, and long-lived biocatalytic activity) [2, 11]. The primary crude extract-based platforms and trends will be discussed.

15.2.1 Platforms

15.2.1.1 Prokaryotic Platforms

E. coli Extract The well-established *E. coli* system provides high protein yields (up to $2.3\,g\,l^{-1}$) [12], as can be seen in Figure 15.3. The system has benefitted from its highly active metabolic activity, as well as the low-cost and scalability of fermentable cells for extract preparation [11]. Notably, the dilute cell-free system has decreased translation elongation rates compared with *in vivo* (~10-fold lower), which improves the expression of mammalian proteins [2]. While perhaps unexpected, it should also be noted that this platform has even had success synthesizing some complex, and even disulfide-bonded proteins [18, 19]. Additionally, well-developed genetic tools to make modifications to the source strain have been critical for developing synthetic genomes that upon cell lysis lead to improved protein production capabilities by removing negative effectors [20]. So far, a limitation of this system is its inability to produce PTMs, such as glycosylation. While PTMs could be enabled through the site-specific introduction of ncAAs (see Section 15.4.1), for example, the inability to introduce PTMs has driven interest in developing eukaryotic platforms.

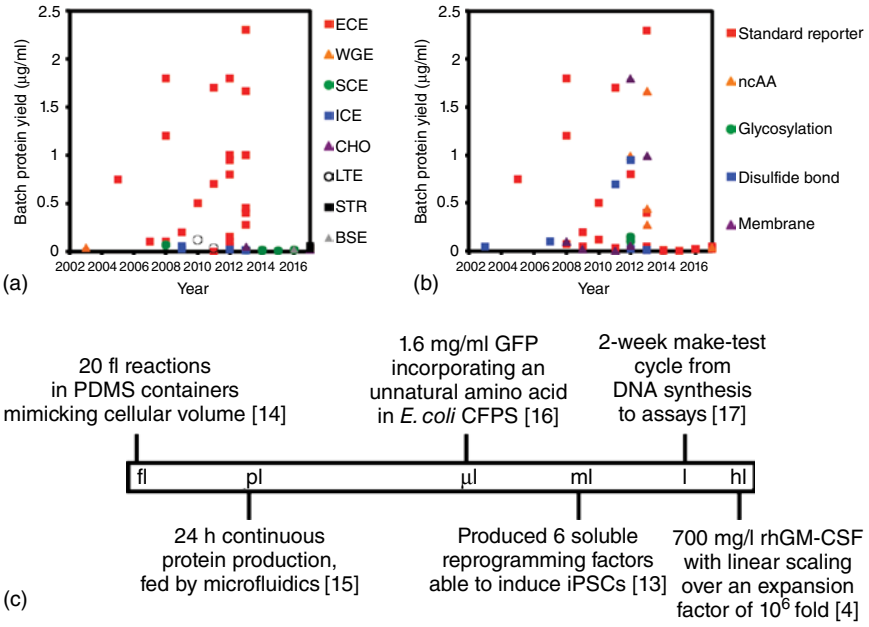


Figure 15.3 Historical trends for different CFPS systems. Batch protein yields for the papers cited in this review are arranged by platform (a) and product type (b). In addition, cell-free protein synthesis has seen successes at a variety of volumes (c). ECE, *E. coli* extract; WGE, wheat germ extract; SCE, *S. cerevisiae* extract; ICE, insect cell extract; CHO, Chinese hamster ovary cell extract; LTE, *L. tarentolae* extract; STR, *Streptomyces* extract; BSE, *B. subtilis* extract; PDMS, polydimethylsiloxane; GFP, green fluorescent protein; iPSCs, induced pluripotent stem cells; rhGM-CSF, recombinant human granulocyte macrophage colony-stimulating factor.

Other Prokaryotic Platforms More recently, alternative prokaryotic platforms have emerged. These platforms have been based on *Bacillus subtilis* [21] and several *Streptomyces* strains: *Streptomyces coelicolor* [22], *Streptomyces lividans* [22], and *Streptomyces venezuelae* [23]. However, the goals around these production systems are more specialized. The *B. subtilis* platform was intended for promoter prototyping and genetic circuits with the hope of translating this to *in vivo* protein expression for metabolic engineering. Alternatively, the *Streptomyces* platform was intended for expression of GC-rich proteins, particularly for expressing and studying natural product gene clusters.

15.2.1.2 Eukaryotic Platforms

In contrast to the *E. coli* CFPS platforms, eukaryotic systems often produce complex proteins with higher percentages of soluble yields. However, they are hampered by comparatively low overall yields (e.g., an order of magnitude in standard batch reactions for similar model proteins) and costly scale-up. While wheat germ extract (WGE) has been the historical eukaryotic system of choice, several promising platforms for industrial use are also now emerging, which include extracts from *Saccharomyces cerevisiae*, insect cells, Chinese hamster ovary (CHO) cells, and *Leishmania tarentolae*, all of which are fermentable, providing possibilities for simple scale-up.

Wheat Germ Extract The WGE system has been the most productive eukaryotic system thus far, producing over 13 000 human proteins in one study [24]. The WGE platform is able to achieve several endogenous PTMs. However, there are aspects of the platform that are not amenable to large-scale protein production. For example, batch yields are typically low (~1–10 $\mu g\,ml^{-1}$ luciferase) [25], the extract preparation is complex, and genetic modifications are challenging. That said, the semicontinuous format has been shown to produce 9.7 $g\,l^{-1}$ green fluorescent protein (GFP) [26]. This is remarkable, enabling the system to be a workhorse for crystallography, NMR, and structural biology studies.

Yeast Extract Pioneered by the work of Iizuka and colleagues, several methods have been used for producing extracts from the yeast *S. cerevisiae*, which is another enticing option for a eukaryotic platform [27]. Like *E. coli*, it is easily grown in a fermenter. Also, the entire genome has been sequenced, and there is a wealth of biological tools, allowing for possible modifications to be made to improve protein production, which was important in the development of the *E. coli* platform.

One method, developed by Wang and colleagues, starts by removing the outer membrane of the cell wall using lyticase, producing a protoplast. Then the protoplast is lysed with a 25-gauge needle. While this method is likely to maintain cellular compartments, the lyticase treatment is expensive on an industrial scale [28].

Other efforts have strived to be more viable as an economical and scalable system. These methods include the use of high-pressure homogenization for cell lysis, combined transcription/translation without need for mRNA capping [29], and a focus on technically simple extract preparation methods [25]. This new method was able to produce $7.69 \pm 0.53\,\mu g\,ml^{-1}$ active luciferase, giving it a fourfold improvement in relative product yield ($\mu g\ \$\ reagent\ cost^{-1}$) over the protoplast method. At this time, it is uncertain whether this approach retains cellular compartments after extract preparation, yet this is a very interesting question. Additionally, using a semicontinuous reaction format to feed limiting substrates (creatine phosphate, nucleotide triphosphates, and perhaps aspartic acid) while removing toxic by-products (inorganic phosphate) led to product yields of $17.0 \pm 3.8\,\mu g\,ml^{-1}$ [30]. Other recent work with the system has explored alternative energy sources [31], fermentation conditions [32], 5′ mRNA leader sequences [33], and gene knockouts [34]. Despite recent work in this system, yields need to be further improved. To do so, a better understanding of the metabolism of the lysate is necessary. Also, elimination of background, nonproductive translation would allow for more efficient use of reactants toward the protein of interest.

Insect Cell Extract Insect cell extract (ICE) systems are another promising platform for eukaryotic CFPS. This approach uses ovary cells of *Spodoptera frugiperda*, the fall armyworm, an industrial *in vivo* protein expression system [35]. Typical yields for the ICE system are ~45 $\mu g\,ml^{-1}$ luciferase [25]. Using mechanical lysis and mild treatment of the extract, a process developed by Kubick and colleagues is able to retain microsomal vesicles of the endoplasmic

reticulum (ER) within the extract [36]. These vesicles are important for trafficking proteins into the ER for membrane insertion and PTMs. The Kubick lab has exploited this by producing membrane proteins, which are able to co-translationally insert into the lipid-enclosed vesicles for stability, as well as glycosylated proteins, both of which will be described later [36, 37]. In addition to glycosylated and membrane proteins, the ICE system has also been demonstrated to incorporate ncAAs using a plasmid developed by the Schultz lab for use in *S. cerevisiae* [38].

Chinese Hamster Ovary Cell Extract CHO cells are widely used industrially for the expression of human recombinant proteins [38]. A benefit is their ability to achieve mammalian PTMs, which remains a challenge. Using the same extract preparation method as ICE, the Kubick lab has begun to develop a highly efficient and high-yielding CHO cell extract. To achieve glycosylation and produce membrane proteins, the reaction mixture can be enriched with microsomal vesicles, yielding 30–50 $\mu g\,ml^{-1}$ of the protein of interest (e.g., luciferase) [38, 39]. This platform offers exciting opportunities for developing advanced process development pipelines for discovering and assaying protein therapeutics, which can be directly translated *in vivo*.

Leishmania tarentolae *Extract* *L. tarentolae*, a lizard parasite, is a fermentable protozoan that was chosen for CFPS. The *in vivo* expression system is able to produce disulfide bonds and glycosylation, and the cells are easy to genetically modify [40, 41]. For extract preparation, a nitrogen cavitation method is used for lysis [42]. A key for the system is that the native mRNA all has the same "splice leader" sequence, allowing for inhibition of endogenous mRNA using an oligonucleotide [40]. This prevents background translation, allowing resources to be directed to synthesis of the protein of interest, producing 50 $\mu g\,ml^{-1}$ GFP. Using the *L. tarentolae* platform, Mureev and colleagues were able to develop species-independent translational sequences (SITS), which allowed for translation in not only *L. tarentolae* platform but also *E. coli* and several eukaryotic cell-free platforms, presumably by a cap-independent pathway [40]. It is expected that this system will aid in expressing proteins from parasitic genomes to test their functions and annotate parasitic genomes, including that of *L. tarentolae* [43].

15.2.2 Trends

Several trends can be observed in the development of the aforementioned cell-free platforms. First, the recent development of several eukaryotic CFPS platforms highlights the enthusiasm and growth of the field.

Second, yields continue to increase for CFPS, with a majority of products expressed in the *E. coli* platform as seen in Figure 15.3a, which catalogs the proteins expressed from manuscripts covered by this review. These improvements have occurred as a result of improved soluble yields for the *E. coli* platform and increased overall yields for the eukaryotic platforms. One method that has been useful in the *E. coli* system was the use of fusion partners to aid

aggregation-prone proteins [13, 44]. Also, moving from glucose to starch as an inexpensive energy source allowed for better pH maintenance, increasing soluble enhanced GFP from 10% to 25% in a study by Kim and colleagues [45], as well as by Caschera and Noireaux [12]. The manuscript by Caschera and Noireaux achieved the highest batch CFPS yield to date or 2.3 g l^{-1} superfolder GFP. The increased yields and decreased cost have enabled the use of freeze-dried lysates for solving cold chain problems with on-demand synthesis of proteins for therapeutics [46, 47] and diagnostics [48, 49]. In contrast to prokaryotic systems, eukaryotic systems generally produce a higher soluble portion but are working toward increasing overall yields cost effectively. So far, this has typically involved reducing background translation, although there are many exciting opportunities for strain engineering. A target goal in the upcoming years is to enable eukaryotic batch CFPS yields of greater than 0.5 mg ml^{-1}, which is chosen because it is about an order of magnitude higher than current levels.

Third, there is also an effort to reduce cost for CFPS. This has been done by moving toward lower cost energy sources, as well as streamlining the process. Instead of fueling the reactions with substrates containing high-energy phosphate bond donors, such as creatine phosphate or phosphoenolpyruvate, *E. coli* reactions have been shown to use glucose and starch as well as nucleoside monophosphates in lieu of triphosphates, greatly reducing cost [12, 45, 50]. So far, eukaryotic systems have not been able to activate cost-effective energy metabolism from non-phosphorylated energy substrates, which will be critical for any industrial-scale applications. Toward more robust and consistent extract preparation methods, extract protocols have been streamlined [51–53]. Another method has combined the small molecules in the reaction into a premix, used T7 polymerase from a crude lysate without purification, and reduced extract preparation by two steps [54].

Fourth, over the last decade, efforts to synthesize complex proteins have intensified. Figure 15.3b, which organizes the values from Figure 15.3a by product, highlights the shift from production of standard reporter proteins, such as luciferase and GFP, toward products containing ncAAs, glycosylation, and disulfide bonds as well as membrane proteins. We expect this trend to continue, particularly given the freedom of design in adjusting cell-free components by the direct addition of new components.

Finally, we note that cell-free platforms have been able to span 17 orders of magnitude in terms of reaction volumes (Figure 15.3c). Notably, the *E. coli* system has been shown to scale linearly from 250 µl reactions to 100 l, an expansion factor of 10^6, producing 700 mg ml^{-1} to enable manufacturing scale synthesis of soluble human granulocyte macrophage colony-stimulating factor (GM-CSF) with two disulfide bonds [4]. In the other direction, there has recently been a move toward smaller, microbe-mimicking reaction sizes [14, 15]. These efforts are useful for high-throughput applications and breadboarding of genetic circuits, both of which will be described later. To learn more about economical scale-up of cell-free systems, see reviews by Swartz [2] and by Carlson *et al.* [6].

The improvements in yields and cost, as well as scalability, give CFPS great utility. Examples of its applications are highlighted in the next section.

15.3 Products

CFPS allows the opportunity to not only produce proteins that standard methods are able to produce but to also solve expression problems with proteins that are notoriously difficult to synthesize *in vivo*. Examples of such products are described in the following section.

15.3.1 Noncanonical Amino Acids

Site-specific incorporation of ncAAs into proteins opens many doors for the production of proteins with new structures, functions, and properties. For such applications, cell-free systems have an advantage over *in vivo* systems because of their open environment and lack of need for cell viability. Indeed, recent efforts by Albayrak and Swartz [55], as well as Jewett and colleagues (unpublished), have shown the ability to synthesize greater yields of protein in batch CFPS reactions as compared with the *in vivo* approach. The benefit appears to come from the fact that the orthogonal translation systems can be toxic to the cell. Moreover, the ncAA can be added directly to the reaction mixture, instead of relying on cellular uptake, and ncAAs can be used that would otherwise be toxic to cells. This technology has been used in cell-free systems to polymerize proteins [16], conjugate human erythropoietin to a fluorophore in ICE [38], and modify the oncoprotein c-Ha-Ras in the WGE [56], along with many others.

The most common method for ncAA incorporation is through amber suppression, which inserts the ncAA at the location of the amber stop codon (UAG) in the reading frame of the gene of interest. With the addition of an orthogonal tRNA, orthogonal aminoacyl-synthetase, and ncAA, the UAG can be incorporated at a specific location in the gene, allowing for the template-encoded addition of the ncAA, as seen in Figure 15.4. This method has been extended to insert a second amino acid using the ochre stop codon (UAA) in combination with the amber codon for the incorporation of two unique ncAAs in a CFPS reaction [57]. Recent advancements from Albayrak and colleagues allow for the synthesis for the orthogonal tRNA (o-tRNA) during the protein synthesis reaction, improving scale-up possibilities [55]. One problem that plagues amber suppression both *in vivo* and *in vitro* is competition between the o-tRNA and release factor 1 (RF1). One solution to this problem is to use a different system for incorporation, using a four-nucleotide codon [58]. Further, cell-free systems open the possibility of expanding the genetic code by introducing additional Watson–Crick base pairs [59] and hijacking sense codons [60]. Since cell viability is no longer an issue, other options remove the problem with RF1 by either adding an aptamer to inhibit it [58] or tagging RF1 and removing it prior to protein synthesis [58, 61]. Looking forward, the development of an RF1 deletion strain as a chassis for CFPS will open new avenues for using cell-free synthetic biology for synthetic chemistry [62].

15.3.2 Glycosylation

For any protein synthesis technology, glycosylation cannot be ignored. It is estimated that over 50% of human proteins are glycosylated [63]. For pure chemical

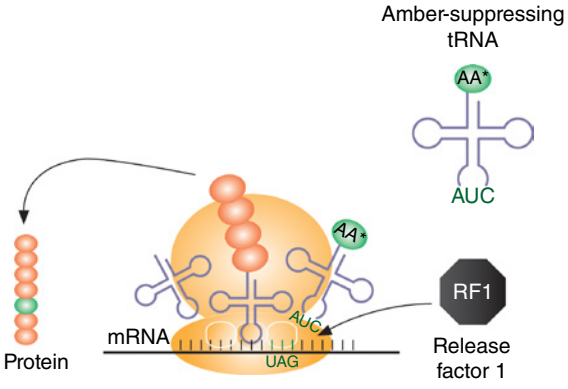


Figure 15.4 The production of proteins containing ncAAs is a frontier of CFPS. Amber suppression, shown here, is the most common method for ncAA incorporation in CFPS platforms but is hampered by competition between the amber-suppressing tRNA and release factor 1 (RF1). Several methods have been developed to prevent this competition. Also, new strains lacking RF1 should address this issue.

synthesis, the stereochemistry of sugars is challenging to make consistently [64], and for *in vivo* protein production, one must use mammalian cells, which are significantly more challenging and more expensive to culture than *E. coli*. This motivates a need for a fast, accurate method for producing glycoproteins using CFPS systems.

Initial work on the production of glycoproteins in CFPS was reported in 1978 by adding canine pancreas microsomes, containing glycosylation machinery, to a WGE reaction [65]. More recently, Guarino and colleagues chose to use the *E. coli* cell-free platform for synthesizing glycoproteins by adding the *Campylobacter jejuni* glycosylation machinery [66]. Since *E. coli* has no native glycosylation machinery, there was no mixture of glycosylation products. Also, due to the open environment of the system, the substrates could be directly added to the reaction to achieve *N*-linked glycosylation. Alternatively, the ICE system is able to maintain microsomes due to the method of lysate production [36]. These microsomes allow for *N*-linked glycosylation, as well as aid in the production of membrane proteins, described later. The CHO cell system had similar results to the ICE system [39]. While efforts to make glycoproteins are underway, there are still two drawbacks: no system is yet able to produce human glycosylation patterns and efforts to achieve O-linked glycosylation are limited. Addressing these limitations will open new avenues for studying and engineering glycosylation. For example, our ability to study and control glycosylation outside the restrictive confines of a cell will help answer fundamental questions such as how glycan attachment affects protein folding and stability. Answers to these questions could lead to general rules for predicting the structural consequences of site-specific protein glycosylation and, in turn, rules for designing modified proteins with advantageous properties.

15.3.3 Antibodies

Antibodies and their variants, typically tackled by *in vivo* recombinant protein methods, have recently gained much attention largely due to their high specificity [67]. However, *in vivo* methods, particularly in prokaryotic cells, can be a challenge when producing high concentrations of antibodies due to their aggregation, leading to insolubility [68]. Yin and colleagues faced this challenge when producing full-length antibodies in the *E. coli* extract (ECE) platform. Notably, they observed that the heavy chain (HC) was more prone to aggregation and needed the light chain (LC) for soluble co-expression [17]. This was an easy problem to solve with the open reaction environment of CFPS. They first expressed the LC plasmid for 1 h and then added the plasmid for the HC to start its translation. This strategy produced 300 mg l^{-1} aglycosylated trastuzumab in reactions ranging from 60 µl to 4 l at greater than 95% solubility. Martin *et al.* were able to then translate this lesson in plasmid timing, as well as oxidizing conditions and chaperone addition, to the CHO CFPS platform for the expression of >100 mg l^{-1} active, intact mAb [69]. In addition to the full-length antibody, antigen-binding fragments [19] and single-chain variable fragments [18, 70, 71] have been produced in a variety of cell-free systems. In fact, notable work by Kanter and colleagues created fusion proteins of a tumor-derived scFv with GM-CSF (a cytokine) or nine amino acids from interleukin-1ß, which improved potency of the scFv by increasing immune system stimulation for cancer therapy [18]. These advances demonstrate the merits of CFPS systems as a potentially powerful antibody production technology. However, cell-free antibody production still struggles from a lack of human glycosylation, which could be achievable in the future through the aforementioned glycosylation methods or ncAA incorporation and coupling of the oligosaccharides.

15.3.4 Membrane Proteins

Membrane proteins are an excellent application for CFPS. Chemical synthesis of membrane proteins can take 1–2 weeks [72], while *in vivo* methods struggle with obtaining high yields, minimizing degradation, and maintaining cell viability [73]. Cell-free systems speed up the process to a matter of hours with decreased proteolysis and no need to maintain living cells. Indeed, CFPS of membrane proteins has received considerable attention in recent years. For example, it has aided in the determination of protein structures, via NMR and crystallography, which were previously impossible, such as ATP synthase and G protein-coupled receptors (GPCRs) [74–76]. The challenge is finding a suitable substitute for the lipid bilayer. As seen in Figure 15.5, these substitutions include the use of detergents (in micelles or bicelles) [74, 77, 78], liposomes [74, 75, 77, 78], nanodiscs [76, 79, 80], tethered bilayer lipid membranes (tBLMs) [81, 82], and microsomal vesicles [36, 37].

One option is to produce the protein, precipitate it, and then solubilize it in detergents or liposomes; however, this does not allow for ideal structure and function studies because it is not an accurate membrane mimic [83]. Further, some detergents cannot be added to the reaction in high concentrations

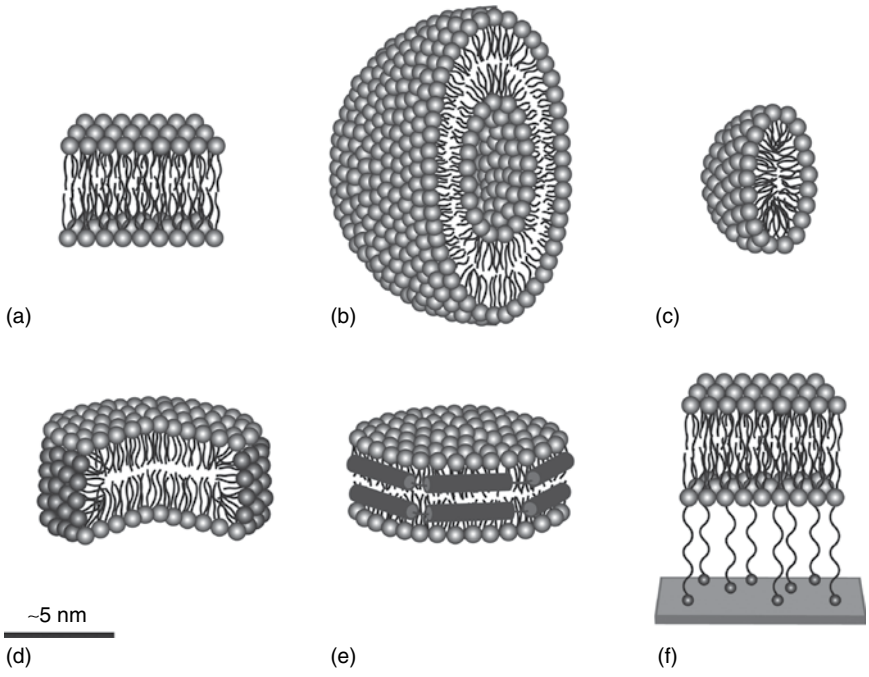


Figure 15.5 CFPS is a useful approach for the production of membrane proteins. Several methods have been implemented to mimic the cell membrane in cell-free protein synthesis: (a) lipid bilayer, (b) liposome, (c) micelle, (d) bicelle, (e) nanodisc, and (f) tethered bilayer lipid membrane.

because they inhibit transcription and translation [75, 83]. These methods also do not take advantage of the open reaction environment of cell-free systems. Unlike cells, where it is impossible to add chemicals directly to the protein as it is synthesized, CFPS allows for co-translation into liposomes, nanodiscs, tBLMs, or microsomes, all of which can be added exogenously to the reaction. Nanodiscs, consisting of a lipid bilayer surrounded by a protein scaffold, were found to be a better mimic of the lipid bilayer and thus obtained higher yields of soluble membrane proteins when compared with detergents and liposomes [80]. In fact, a functional GPCR, a highly studied but difficult to produce protein, was first produced in soluble form using nanodiscs in a cell-free reaction [76]. Another useful aspect of nanodiscs is the ability to co-express the nanodisc protein scaffold and membrane protein in the cell-free reaction, reducing the number of production and purification steps necessary [79]. For deeper structural and functional studies, the tBLMs use self-assembly to attach a membrane to a gold surface. The protein can then be co-translationally inserted into the membrane and immediately studied using surface plasmon-enhanced fluorescence spectroscopy (SPFS) and imaging surface plasmon resonance (iSPR), fluorescence polarization (FP) [84]. Similar to the tBLMs, CFPS has also been used in conjunction with a phospholipid bilayer supported on quartz crystal microbalances for direct characterization of membrane proteins as they

are expressed [85]. CFPS of membrane proteins promises to help unravel the function and structure of many potential drug targets.

15.4 High-Throughput Applications

Processes that take days or weeks to design, prepare, and execute *in vivo* can often be done more rapidly in a cell-free system. The use of polymerase chain reaction (PCR) templates significantly speeds up the process, since no time-consuming cloning steps are needed. Also, since the cell-free system is simpler and easier to control than cells, it allows for direct manipulation of reaction environments, as well as optimization of the reaction conditions. These characteristics are highlighted in the following examples of high-throughput protein synthesis for both production and screening as well as genetic circuit designing and testing.

15.4.1 Protein Production and Screening

While chemistry has been able to produce small molecule libraries for easy screening, the ability to produce proteins for similar procedures has been challenging. However, with cell-free systems, there is no need to transform cells with plasmids, produce the protein, and then lyse the cells. Instead, a PCR template or plasmid can be added to a small reaction mixture in a plate, the protein can be produced, and then the various proteins on the plate can be screened *in situ*, all in a matter of hours [86]. For example, Karim and Jewett expressed several enzymes in a CFPS reaction for prototyping metabolic pathways in *E. coli* lysates in order to quickly arrive upon the best combination of enzymes for the production of butanol [87]. Since CFPS reactions are at a small scale, microfluidics can also be used to supply small molecules [88] or when the number of reactions becomes too large, liquid handling can easily be automated [89]. One of the most impressive examples of using CFPS for high-throughput protein production is the human protein factory [24]. In this study, the authors expressed 13,364 human proteins using the WGE platform and then compiled the protein expression information in an online database [24, 90].

In addition to producing proteins from standard plasmids and PCR products, it is possible to produce protein arrays from DNA arrays. Since DNA arrays are much easier and more stable than protein arrays, He and colleagues developed a method to "stamp" the proteins on a new array by putting a DNA array plate face down on a second plate with the CFPS reaction mixture between the plates [91]. After the proteins were produced, they associated with the surface of the new plate. Stoevesandt and colleagues demonstrated the utility of this method when they produced an array of 116 distinct proteins [92]. In addition to its ease, it was found that one DNA array was able to produce at least 20 new protein arrays [91]. Protein arrays are beginning to enable an improved toolbox, and a faster process to probe different aspects of protein function and their role in enzyme screening will continue to grow in the upcoming years.

15.4.2 Genetic Circuit Optimization

There is currently a need for "breadboarding" of *in vivo* biological circuits in order to accelerate the design–build–test loops associated with synthetic biology studies. Biological circuits rely on regulation and control of protein products and can take a long time to assemble *in vivo*, so a system is needed that will function similarly to the cell with faster results and greater flexibility for manipulation: a great application for CFPS platforms. The combinatorial nature of testing the variations of the circuits also lends itself to high-throughput methods. Also, the open environment of the CFPS reaction allows for more control for these studies, since the initial concentrations of mRNA and protein as well as the exact reaction size can be directly manipulated. Methods have been developed to characterize parts (e.g., promoters, ribosome binding sites, terminators, and spacing), as well as multienzyme systems, such that they function predictably both *in vitro* and *in vivo* [21, 23, 93–96]. In one such example, Chappell and colleagues recognized that ribosome binding sites correlated directly when using PCR products *in vitro*, but promoters did not [94]. Thus, they used a USER–ligase method to circularize PCR products, the results of which were able to correlate between both platforms while keeping production time short by avoiding the need for a plasmid typically obtained by cell growth. In addition to characterization, cell-free systems have been used to test new options for circuit proteins, such as endogenous sigma factors, to supplement the common LacI and TetR proteins [7]. Aiding in the high-throughput area, reactions at the nanoliter, picoliter, and femtoliter scales are being explored as a method to better approximate the volume of a cell. This involves using microfluidics to feed small molecules to the reaction [15, 97], which diffuse well due to the small volume, as well as studying noise in gene expression [14], which could aid in the future design of gene circuits. To learn more about *in vitro* genetic circuits, see a review by Hockenberry [98].

15.5 Future of the Field

CFPS is emerging as a disruptive technology. It has promising applications for rapid, high-throughput screening and production of enzymes and personalized medicines, membrane proteins, and proteins containing ncAAs. Other applications include efforts to construct fully synthetic ribosomes *in vitro* [99] as well as artificial cells [7, 100]. Equally important, CFPS is expected to help address the increasing discrepancy between genome sequence data and their translation products. The Sargasso Sea expedition alone, for example, generated 1.2 million new genes, many with unknown function [101]. This concept has already been proven by the expression of the entire T7 bacteriophage genome [102] as well as nanoassemblies of T4 bacteriophage structural proteins [103]. Unfortunately, current cell-based technologies for heterologous protein expression have been unable to meet the rapidly expanding need for affordable, simple, and efficient protein production because they (i) can be slow (requiring time-consuming cloning strategies), (ii) can require laborious protein purification procedures,

and (iii) can lack robustness and predictability due to several reasons: the complexity, the host-dependent gene expression and protein folding/function, the necessity of product export from the cell membrane for improved production, and the toxicity of high levels of expressed proteins to the host. CFPS can address many of these limitations to help complement existing technologies, but there are remaining immediate challenges. For example, the field is limited by its ability to produce posttranslationally modified proteins at high titers, particularly those with human patterns. Moreover, we still do not have the protein equivalent of PCR. Further, inefficiencies in site-specific incorporation of ncAAs limits innovation. By addressing these challenges, we anticipate that cell-free systems will continue to penetrate and be recognized for value by industry. Given the capability to modify and control cell-free systems, CFPS holds promise to be a powerful tool for systems biology, for synthetic biology, and as a protein production technology in years to come.

Definitions

Cell-free protein synthesis is the process of translating proteins in lysates
In vitro is the processes performed outside of their biological context, e.g. protein synthesis occurring outside the cell
Noncanonical amino acid is any amino acid outside the 20 canonical amino acids
Glycosylation is the addition of sugar moieties to proteins
Antibody is the protein of the immune system that recognizes and neutralizes pathogens
Membrane protein is the protein that is associated with or integrated into a cellular membrane
High-throughput is the capability of being performed many times in parallel
Protein screening is the process of testing one or more proteins or protein variants in one or more contexts to determine properties of the protein(s) or optimize
Genetic circuit is the engineered use of DNA sequences to control biological reactions and programs

Acknowledgments

The authors thank C. Eric Hodgman for his advice and discussions regarding the manuscript. MCJ gratefully acknowledges funding from the National Science Foundation (Grant Number MCB-0943393, Grant Number DMR-1108350, Grant Number MCB-1716766), the Air Force Research laboratory (FA8650-15-2-5518), the Army Research Office (W911NF-11-1-0445 and W911NF-16-1-0372), the Human Frontiers Science Program (Grant Number RGP0015/2017), the Office of Naval Research (Grant Number N00014-11-1-0363), the DARPA Living Foundries Program (N66001-12-C-4211), the DARPA Biomedicines on Demand Program (N66001-13-C-4024), the Defense Threat Reduction Agency (GRANT11631647), the Department of Energy (Grant Number DE-SC0018249), the David and Lucile

Packard Foundation, the Dreyfus Teacher-Scholar Program, and the Chicago Biomedical Consortium with support from the Searle Funds at the Chicago Community Trust. JAS was supported by the National Science Foundation Graduate Research Fellowship (Grant Number DGE-1324585). The authors declare no commercial or financial conflict of interest.

References

1 Nirenberg, M.W. and Matthaei, J.H. (1961) The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. U.S.A.*, **47**, 1588–1602.
2 Swartz, J.R. (2006) Developing cell-free biology for industrial applications. *J. Ind. Microbiol. Biotechnol.*, **33**, 476–485. doi: 10.1007/s10295-006-0127-y
3 Katzen, F., Chang, G., and Kudlicki, W. (2005) The past, present and future of cell-free protein synthesis. *Trends Biotechnol.*, **23**, 150–156. doi: 10.1016/j.tibtech.2005.01.003
4 Zawada, J.F., Yin, G., Steiner, A.R., Yang, J., Naresh, A., Roy, S.M., Gold, D.S., Heinsohn, H.G., and Murray, C.J. (2011) Microscale to manufacturing scale-up of cell-free cytokine production – a new approach for shortening protein production development timelines. *Biotechnol. Bioeng.*, **108**, 1570–1578. doi: 10.1002/bit.23103
5 Ranji, A., Wu, J.C., Bundy, B.C., and Jewett, M.C. (2013) Transforming synthetic biology with cell-free systems, in *Synthetic Biology Tools and Applications* (ed. H. Zhao), Elsevier, New York, pp. 277–302.
6 Carlson, E.D., Gan, R., Hodgman, C.E., and Jewett, M.C. (2012) Cell-free protein synthesis: applications come of age. *Biotechnol. Adv.*, **30**, 1185–1194. doi: 10.1016/j.biotechadv.2011.09.016
7 Shin, J. and Noireaux, V. (2012) An *E. coli* cell-free expression toolbox: application to synthetic gene circuits and artificial cells. *ACS Synth. Biol.*, **1**, 29–41. doi: 10.1021/sb200016s
8 Ohashi, H., Kanamori, T., Shimizu, Y., and Ueda, T. (2010) A highly controllable reconstituted cell-free system – a breakthrough in protein synthesis research. *Curr. Pharm. Biotechnol.*, **11**, 267–271. doi: 10.2174/138920110791111889
9 Hillebrecht, J.R. and Chong, S. (2008) A comparative study of protein synthesis in in vitro systems: from the prokaryotic reconstituted to the eukaryotic extract-based. *BMC Biotechnol.*, **8**, 58. doi: 10.1186/1472-6750-8-58
10 Asahara, H. and Chong, S. (2010) In vitro genetic reconstruction of bacterial transcription initiation by coupled synthesis and detection of RNA polymerase holoenzyme. *Nucleic Acids Res.*, **38**, e141. doi: 10.1093/nar/gkq377
11 Swartz, J.R. (2012) Transforming biochemical engineering with cell-free biology. *AIChE J.*, **58**, 5–13. doi: 10.1002/aic.13701
12 Caschera, F. and Noireaux, V. (2014) Synthesis of 2.3 mg/ml of protein with an all *Escherichia coli* cell-free transcription-translation system. *Biochimie*, **99**, 162–168. doi: 10.1016/j.biochi.2013.11.025
13 Yang, W.C., Patel, K.G., Lee, J., Ghebremariam, Y.T., Wong, H.E., Cooke, J.P., and Swartz, J.R. (2009) Cell-free production of transducible transcription factors for nuclear reprogramming. *Biotechnol. Bioeng.*, **104**, 1047–1058. doi: 10.1002/bit.22517

14 Karig, D.K., Jung, S.Y., Srijanto, B., Collier, C.P., and Simpson, M.L. (2013) Probing cell-free gene expression noise in femtoliter volumes. *ACS Synth. Biol.* doi: 10.1021/sb400028c

15 Siuti, P., Retterer, S.T., and Doktycz, M.J. (2011) Continuous protein production in nanoporous, picolitre volume containers. *Lab Chip*, **11**, 3523–3529. doi: 10.1039/c1lc20462a

16 Albayrak, C. and Swartz, J.R. (2013) Direct polymerization of proteins. *ACS Synth. Biol.* doi: 10.1021/sb400116x

17 Yin, G., Garces, E.D., Yang, J., Zhang, J. *et al.* (2012) Aglycosylated antibodies and antibody fragments produced in a scalable in vitro transcription-translation system. *MAbs*, **4**, 217–225. doi: 10.4161/mabs.4.2.19202

18 Kanter, G., Yang, J., Voloshin, A., Levy, S., Swartz, J.R., and Levy, R. (2007) Cell-free production of scFv fusion proteins: an efficient approach for personalized lymphoma vaccines. *Blood*, **109**, 3393–3399. doi: 10.1182/blood-2006-07-030593

19 Oh, I.S., Lee, J.C., Lee, M.S., Chung, J.H. *et al.* (2010) Cell-free production of functional antibody fragments. *Bioprocess. Biosyst. Eng.*, **33**, 127–132. doi: 10.1007/s00449-009-0372-3

20 Michel-Reydellet, N., Calhoun, K.A., and Swartz, J.R. (2004) Amino acid stabilization for cell-free protein synthesis by modification of the *Escherichia coli* genome. *Metab. Eng.*, **6**, 197–203.

21 Kelwick, R., Webb, A.J., MacDonald, J.T., and Freemont, P.S. (2016) Development of a *Bacillus subtilis* cell-free transcription-translation system for prototyping regulatory elements. *Metab. Eng.*, **38**, 370–381. doi: 10.1016/j.ymben.2016.09.008

22 Li, J., Wang, H., Kwon, Y.-C., and Jewett, M.C. (2017) Establishing a high yielding streptomyces-based cell-free protein synthesis system. *Biotechnol. Bioeng.*, **114**, 1343–1353.

23 Moore, S.J., Lai, H.E., Needham, H., Polizzi, K.M., and Freemont, P.S. (2017) Streptomyces venezuelae TX-TL – a next generation cell-free synthetic biology tool. *Biotechnol. J.*, **12**, 1–7. doi: 10.1002/biot.201600678

24 Goshima, N., Kawamura, Y., Fukumoto, A., Miura, A., Honma, R., Satoh, R., Wakamatsu, A., Yamamoto, J., Kimura, K., Nishikawa, T., Andoh, T., Iida, Y., Ishikawa, K., Ito, E., Kagawa, N., Kaminaga, C., Kanehori, K., Kawakami, B., Kenmochi, K., Kimura, R., Kobayashi, M., Kuroita, T., Kuwayama, H., Maruyama, Y., Matsuo, K., Minami, K., Mitsubori, M., Mori, M., Morishita, R., Murase, A., Nishikawa, A., Nishikawa, S., Okamoto, T., Sakagami, N., Sakamoto, Y., Sasaki, Y., Seki, T., Sono, S., Sugiyama, A., Sumiya, T., Takayama, T., Takayama, Y., Takeda, H., Togashi, T., Yahata, K., Yamada, H., Yanagisawa, Y., Endo, Y., Imamoto, F., Kisu, Y., Tanaka, S., Isogai, T., Imai, J., Watanabe, S., and Nomura, N. (2008) Human protein factory for converting the transcriptome into an in vitro–expressed proteome. *Nat. Methods*, **5**, 1011–1017. doi: 10.1038/nmeth.1273

25 Hodgman, C.E. and Jewett, M.C. (2013) Optimized extract preparation methods and reaction conditions for improved yeast cell-free protein synthesis. *Biotechnol. Bioeng.*, **110**, 2643–2654. doi: 10.1002/bit.24942/abstract

26 Endo, Y. and Sawasaki, T. (2006) Cell-free expression systems for eukaryotic protein production. *Curr. Opin. Biotechnol.*, **17**, 373–380. doi: 10.1016/j.copbio.2006.06.009

27 Iizuka, N., Najita, L., Franzusoff, A., and Sarnow, P. (1994) Cap-dependent and cap-independent translation by internal initiation of mRNAs in cell extracts prepared from *Saccharomyces cerevisiae*. *Mol. Cell. Biol.*, **14**, 7322–7330. doi: 10.1128/mcb.14.11.7322

28 Wang, X., Liu, J., Zheng, Y., Li, J. *et al.* (2008) An optimized yeast cell-free system: sufficient for translation of human papillomavirus 58 L1 mRNA and assembly of virus-like particles. *J. Biosci. Bioeng.*, **106**, 8–15. doi: 10.1263/jbb.106.8

29 Gan, R. and Jewett, M.C. (2014) A combined cell-free transcription-translation system from *Saccharomyces cerevisiae* for rapid and robust protein synthe. *Biotechnol. J.*, **9**, 641–651. doi: 10.1002/biot.201300545

30 Schoborg, J.A., Hodgman, C.E., Anderson, M.J., and Jewett, M.C. (2014) Substrate replenishment and byproduct removal improve yeast cell-free protein synthesis. *Biotechnol. J.*, **9**, 630–640.

31 Anderson, M.J., Stark, J.C., Hodgman, C.E., and Jewett, M.C. (2015) Energizing eukaryotic cell-free protein synthesis with glucose metabolism. *FEBS Lett.*, **589**, 1723–1727. doi: 10.1016/j.febslet.2015.05.045

32 Choudhury, A., Hodgman, C.E., Anderson, M.J., and Jewett, M.C. (2014) Evaluating fermentation effects on cell growth and crude extract metabolic activity for improved yeast cell-free protein synthesis. *Biochem. Eng. J.*, **91**, 140–148. doi: 10.1016/j.bej.2014.07.014

33 Hodgman, C.E. and Jewett, M.C. (2014) Characterizing IGR IRES-mediated translation initiation for use in yeast cell-free protein synthesis. *New Biotechnol.*, **31**, 499–505. doi: 10.1016/j.nbt.2014.07.001

34 Schoborg, J.A., Clark, L.G., Choudhury, A., Hodgman, C.E., and Jewett, M.C. (2016) Yeast knockout library allows for efficient testing of genomic mutations for cell-free protein synthesis. *Synth. Syst. Biotechnol.*, 1–4. doi: 10.1016/j.synbio.2016.02.004

35 Drugmand, J.C., Schneider, Y.J., and Agathos, S.N. (2012) Insect cells as factories for biomanufacturing. *Biotechnol. Adv.*, **30**, 1140–1157. doi: 10.1016/j.biotechadv.2011.09.014

36 Kubick, S., Gerrits, M., Merk, H., Stiege, W. *et al.* (2009) In vitro synthesis of posttranslationally modified membrane proteins, in *Current Topics in Membranes* (eds D.J. Benos and S.A. Simon), Elsevier, pp. 25–49. doi: 10.1016/s1063-5823(09)63002-7

37 Sachse, R., Wüstenhagen, D., Šamalíková, M., Gerrits, M. *et al.* (2013) Synthesis of membrane proteins in eukaryotic cell-free systems. *Eng. Life Sci.*, **13**, 39–48. doi: 10.1002/elsc.201100235

38 Stech, M., Brodel, A.K., Quast, R.B., Sachse, R. *et al.* (2013) *Cell-free systems: functional modules for synthetic and chemical biology*, Adv. Biochem. Eng, Biotechnol. doi: 10.1007/10_2013_185

39 Brodel, A.K., Sonnabend, A., and Kubick, S. (2014) Cell-free protein expression based on extracts from CHO cells. *Biotechnol. Bioeng.*, **111**, 25–36. doi: 10.1002/bit.25013/abstract

40 Mureev, S., Kovtun, O., Nguyen, U.T., and Alexandrov, K. (2009) Species-independent translational leaders facilitate cell-free expression. *Nat. Biotechnol.*, **27**, 747–752. doi: 10.1038/nbt.1556

41 Breitling, R., Klingner, S., Callewaert, N., Pietrucha, R. *et al.* (2002) Non-pathogenic trypanosomatid protozoa as a platform for protein research and production. *Protein Expression Purif.*, **25**, 209–218.

42 Kovtun, O., Mureev, S., Jung, W., Kubala, M.H., Johnston, W., and Alexandrov, K. (2011) Leishmania cell-free protein expression system. *Methods*, **55**, 58–64. doi: 10.1016/j.ymeth.2011.06.006

43 Kovtun, O., Mureev, S., Johnston, W., and Alexandrov, K. (2010) Towards the construction of expressed proteomes using a *Leishmania tarentolae* based cell-free expression system. *PLoS One*, **5**, e14388. doi: 10.1371/journal.pone.0014388.g001

44 Ahn, J.H., Keum, J.W., and Kim, D.M. (2011) Expression screening of fusion partners from an *E. coli* genome for soluble expression of recombinant proteins in a cell-free protein synthesis system. *PLoS One*, **6**, e26875. doi: 10.1371/journal.pone.0026875

45 Kim, H.-C., Kim, T.-W., and Kim, D.-M. (2011) Prolonged production of proteins in a cell-free protein synthesis system using polymeric carbohydrates as an energy source. *Process Biochem.*, **46**, 1366–1369. doi: 10.1016/j.procbio.2011.03.008

46 Karig, D.K., Bessling, S., Thielen, P., Zhang, S., and Wolfe, J. (2016) Preservation of protein expression systems at elevated temperatures for portable therapeutic production. *J. R. Soc. Interface*, **14**, 20161039.

47 Pardee, K., Slomovic, S., Nguyen, P.Q., Lee, J.W., Donghia, N., Burrill, D., Ferrante, T., McSorley, F.R., Furuta, Y., Vernet, A., Lewandowski, M., Boddy, C.N., Joshi, N.S., and Collins, J.J. (2016) Portable, on-demand biomolecular manufacturing. *Cell*, **167**, 248–259.e12. doi: 10.1016/j.cell.2016.09.013

48 Pardee, K., Green, A.A., Ferrante, T., Cameron, D.E., Daley Keyser, A., Yin, P., and Collins, J.J. (2014) Paper-based synthetic gene networks. *Cell*, **159**, 940–954. doi: 10.1016/j.cell.2014.10.004

49 Pardee, K., Green, A.A., Takahashi, M.K., Braff, D., Lambert, G., Lee, J.W., Ferrante, T., Ma, D., Donghia, N., Fan, M., Daringer, N.M., Bosch, I., Dudley, D.M., O'Connor, D.H., Gehrke, L., and Collins, J.J. (2016) Rapid, low-cost detection of zika virus using programmable biomolecular components. *Cell*, **165**, 1255–1266. doi: 10.1016/j.cell.2016.04.059

50 Calhoun, K.A. and Swartz, J.R. (2005) An economical method for cell-free protein synthesis using glucose and nucleoside monophosphates. *Biotechnol. Progr.*, **21**, 1146–1153.

51 Kim, T.W., Keum, J.W., Oh, I.S., Choi, C.Y. *et al.* (2006) Simple procedures for the construction of a robust and cost-effective cell-free protein synthesis system. *J. Biotechnol.*, **126**, 554–561. doi: 10.1016/j.jbiotec.2006.05.014

52 Kim, T.-W., Kim, H.-C., Oh, I.-S., and Kim, D.-M. (2008) A highly efficient and economical cell-free protein synthesis system using the S12 extract of *Escherichia coli*. *Biotechnol. Bioprocess Eng.*, **13**, 464–469. doi: 10.1007/s12257-008-0139-8

53 Shrestha, P., Holland, T.M., and Bundy, B.C. (2012) Streamlined extract preparation for *Escherichia coli*-based cell-free protein synthesis by sonication or bead vortex mixing. *Biotechniques*, **53**, 163–174. doi: 10.2144/0000113924

54 Yang, W.C., Patel, K.G., Wong, H.E., and Swartz, J.R. (2012) Simplifying and streamlining *Escherichia coli*-based cell-free protein synthesis. *Biotechnol. Progr.*, **28**, 413–420. doi: 10.1002/btpr.1509

55 Albayrak, C. and Swartz, J.R. (2013) Cell-free co-production of an orthogonal transfer RNA activates efficient site-specific non-natural amino acid incorporation. *Nucleic Acids Res.*, **41**, 5949–5963. doi: 10.1093/nar/gkt226

56 Yabuki, T., Kigawa, T., Dohmae, N., Takio, K. *et al.* (1998) Dual amino acid-selective and site-directed stable-isotope labeling of the human c-Ha-Ras protein by cell-free synthesis. *J. Biomol. NMR*, **11**, 295–306.

57 Ozer, E., Chemla, Y., Schlesinger, O., Aviram, H.Y., Riven, I., Haran, G., and Alfonta, L. (2017) In vitro suppression of two different stop codons. *Biotechnol. Bioeng.*, **114**, 1065–1073. doi: 10.1002/bit.26226

58 Lee, K.B., Kim, H.C., Kim, D.M., Kang, T.J. *et al.* (2012) Comparative evaluation of two cell-free protein synthesis systems derived from *Escherichia coli* for genetic code reprogramming. *J. Biotechnol.*, **164**, 330–335. doi: 10.1016/j.jbiotec.2013.01.011

59 Hirao, I., Ohtsuki, T., Fujiwara, T., Mitsui, T. *et al.* (2002) An unnatural base pair for incorporating amino acid analogs into proteins. *Nat. Biotechnol.*, **20**, 177–182.

60 Goto, Y., Katoh, T., and Suga, H. (2011) Flexizymes for genetic code reprogramming. *Nat. Protoc.*, **6**, 779–790. doi: 10.1038/nprot.2011.331

61 Loscha, K.V., Herlt, A.J., Qi, R., Huber, T. *et al.* (2012) Multiple-site labeling of proteins with unnatural amino acids. *Angew. Chem. Int. Ed.*, **51**, 2243–2246. doi: 10.1002/anie.201108275

62 Hong, S.H., Ntai, I., Haimovich, A.D., Kelleher, N.L. *et al.* (2014) Cell-free protein synthesis from a release factor 1 deficient *Escherichia coli* activates efficient and multiple site-specific nonstandard amino acid incorporation. *ACS Synth. Biol.* doi: 10.1021/sb400140t

63 Apweiler, R., Hermjakob, H., and Sharon, N. (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim. Biophys. Acta*, **1473**, 4–8.

64 Seeberger, P.H. (2008) Automated oligosaccharide synthesis. *Chem. Soc. Rev.*, **37**, 19–28. doi: 10.1039/b511197h

65 Lingappa, V.R., Lingappa, J.R., Prasad, R., Ebner, K. *et al.* (1978) Coupled cell-free synthesis, segregation, and core glycosylation of a secretory protein. *Proc. Natl. Acad. Sci. U.S.A.*, **75**, 2338–2342.

66 Guarino, C. and DeLisa, M.P. (2012) A prokaryote-based cell-free translation system that efficiently synthesizes glycoproteins. *Glycobiology*, **22**, 596–601. doi: 10.1093/glycob/cwr151

67 Sapra, P. and Shor, B. (2013) Monoclonal antibody-based therapies in cancer: advances and challenges. *Pharmacol. Ther.*, **138**, 452–469. doi: 10.1016/j.pharmthera.2013.03.004

68 Baneyx, F. and Mujacic, M. (2004) Recombinant protein folding and misfolding in *Escherichia coli*. *Nat. Biotechnol.*, **22**, 1399–1408. doi: 10.1038/nbt1029

69 Martin, R.W., Majewska, N.I., Chen, C.X., Albanetti, T.E., Jimenez, R.B., Schmelzer, A.E., Jewett, M.C., and Roy, V. (2017) Development of a CHO-based cell-free platform for synthesis of active monoclonal antibodies. *ACS Synth. Biol.* doi: 10.1021/acssynbio.7b00001

70 Stech, M., Merk, H., Schenk, J.A., Stocklein, W.F., Wustenhagen, D.A., Micheel, B., Duschl, C., Bier, F.F., and Kubick, S. (2012) Production of functional antibody

fragments in a vesicle-based eukaryotic cell-free translation system. *J. Biotechnol.*, **164**, 220–231. doi: 10.1016/j.jbiotec.2012.08.020

71 Kawasaki, T., Gouda, M.D., Sawasaki, T., Takai, K., and Endo, Y. (2003) Efficient synthesis of a disulfide-containing protein through a batch cell-free system from wheat germ. *Eur. J. Biochem.*, **270**, 4780–4786. doi: 10.1046/j.1432-1033.2003.03880.x

72 Kochendoerfer, G.G., Salom, D., Lear, J.D., Wilk-Orescan, R. *et al.* (1999) Total chemical synthesis of the integral membrane protein influenza A virus M2: role of Its C-terminal domain in tetramer assembly. *Biochemistry*, **38**, 11905–11913. doi: 10.1021/bi990720m

73 Liguori, L., Marques, B., Villegas-Méndez, A., Rothe, R. *et al.* (2007) Production of membrane proteins using cell-free expression systems. *Expert Rev. Proteomics*, **4**, 79–90. doi: 10.1586/14789450.4.1.79

74 Matthies, D., Haberstock, S., Joos, F., Dotsch, V. *et al.* (2011) Cell-free expression and assembly of ATP synthase. *J. Mol. Biol.*, **413**, 593–603. doi: 10.1016/j.jmb.2011.08.055

75 Uhlemann, E.M., Pierson, H.E., Fillingame, R.H., and Dmitriev, O.Y. (2012) Cell-free synthesis of membrane subunits of ATP synthase in phospholipid bicelles: NMR shows subunit a fold similar to the protein in the cell membrane. *Protein Sci.*, **21**, 279–288. doi: 10.1002/pro.2014

76 Yang, J.P., Cirico, T., Katzen, F., Peterson, T.C. *et al.* (2011) Cell-free synthesis of a functional G protein-coupled receptor complexed with nanometer scale bilayer discs. *BMC Biotechnol.*, **11**, 57. doi: 10.1186/1472-6750-11-57

77 Klammt, C., Lohr, F., Schafer, B., Haase, W. *et al.* (2004) High level cell-free expression and specific labeling of integral membrane proteins. *Eur. J. Biochem.*, **271**, 568–580. doi: 10.1111/j.1432-1033.2003.03959.x

78 Schwarz, D., Junge, F., Durst, F., Frolich, N. *et al.* (2007) Preparative scale expression of membrane proteins in *Escherichia coli*-based continuous exchange cell-free systems. *Nat. Protoc.*, **2**, 2945–2957. doi: 10.1038/nprot.2007.426

79 Cappuccio, J.A., Blanchette, C.D., Sulcheck, T.A., Arroyo, E.S., Kralj, J.M., Hinz, A.K., Kuhn, E.A., Chromy, B.A., Segelke, B.W., Rothschild, K.J., Fletcher, J.E., Katzen, F., Peterson, T.C., Kudlicki, W.A., Bench, G., Hoeprich, P.D., and Coleman, M.A. (2008) Cell-free co-expression of functional membrane proteins and apolipoprotein, forming soluble nanolipoprotein particles. *Mol. Cell. Proteomics*, **7**, 2246–2253. doi: 10.1074/

80 Lyukmanova, E.N., Shenkarev, Z.O., Khabibullina, N.F., Kopeina, G.S. *et al.* (2012) Lipid-protein nanodiscs for cell-free production of integral membrane proteins in a soluble and folded state: comparison with detergent micelles, bicelles and liposomes. *Biochim. Biophys. Acta*, **1818**, 349–358. doi: 10.1016/j.bbamem.2011.10.020

81 Yildiz, A.A., Knoll, W., Gennis, R.B., and Sinner, E.K. (2012) Cell-free synthesis of cytochrome bo(3) ubiquinol oxidase in artificial membranes. *Anal. Biochem.*, **423**, 39–45. doi: 10.1016/j.ab.2012.01.007

82 Damiati, S., Zayni, S., Schrems, A., Kiene, E., Sleytr, U.B., Chopineau, J., Schuster, B., and Sinner, E.-K. (2015) Inspired and stabilized by nature: ribosomal synthesis of the human voltage gated ion channel (VDAC) into 2D-protein-tethered lipid interfaces. *Biomater. Sci.*, **3**, 1406–1413. doi: 10.1039/C5BM00097A

83 Katzen, F., Peterson, T.C., and Kudlicki, W. (2009) Membrane protein expression: no cells required. *Trends Biotechnol.*, **27**, 455–460. doi: 10.1016/j.tibtech.2009.05.005

84 Yildiz, A.A., Kang, C., and Sinner, E.K. (2013) Biomimetic membrane platform containing hERG potassium channel and its application to drug screening. *Analyst*, **138**, 2007–2012. doi: 10.1039/c3an36159d

85 Chalmeau, J., Monina, N., Shin, J., Vieu, C. *et al.* (2011) α-Hemolysin pore formation into a supported phospholipid bilayer using cell-free expression. *Biochim. Biophys. Acta*, **1808**, 271–278. doi: 10.1016/j.bbamem.2010.07.027

86 Kwon, Y.C., Lee, K.H., Kim, H.C., Han, K., Seo, J.H., Kim, B.G., and Kim, D.M. (2010) Cloning-independent expression and analysis of omega-transaminases by use of a cell-free protein synthesis system. *Appl. Environ. Microbiol.*, **76**, 6295–6298. doi: 10.1128/AEM.00029-10

87 Karim, A.S. and Jewett, M.C. (2016) A cell-free framework for rapid biosynthetic pathway prototyping and enzyme discovery. *Metab. Eng.*, **36**, 116–126.

88 Khnouf, R., Olivero, D., Jin, S., Coleman, M.A., and Fan, Z.H. (2010) Cell-free expression of soluble and membrane proteins in an array device for drug screening. *Anal. Chem.*, **82**, 7021–7026. doi: 10.1002/btpr.474

89 Matsuoka, K., Komori, H., Nose, M., Endo, Y. *et al.* (2009) Simple screening method for autoantigen proteins using the N-terminal biotinylated protein library produced by wheat cell-free synthesis. *J. Proteome Res.*, **9**, 4264–4273. doi: 10.1021/pr9010553

90 Maruyama, Y., Wakamatsu, A., Kawamura, Y., Kimura, K. *et al.* (2009) Human gene and protein database (HGPD): a novel database presenting a large quantity of experiment-based results in human proteomics. *Nucleic Acids Res.*, **37**, D762–D766. doi: 10.1093/nar/gkn872

91 He, M., Stoevesandt, O., Palmer, E.A., Khan, F. *et al.* (2008) Printing protein arrays from DNA arrays. *Nat. Methods*, **5**, 175–177. doi: 10.1038/nmeth.1178

92 Stoevesandt, O., Vetter, M., Kastelic, D., Palmer, E.A. *et al.* (2011) Cell free expression put on the spot: advances in repeatable protein arraying from DNA (DAPA). *New Biotechnol.*, **28**, 282–290. doi: 10.1016/j.nbt.2010.09.004

93 Karig, D.K., Iyer, S., Simpson, M.L., and Doktycz, M.J. (2012) Expression optimization and synthetic gene networks in cell-free systems. *Nucleic Acids Res.*, **40**, 3763–3774. doi: 10.1093/nar/gkr1191

94 Chappell, J., Jensen, K., and Freemont, P.S. (2013) Validation of an entirely in vitro approach for rapid prototyping of DNA regulatory elements for synthetic biology. *Nucleic Acids Res.*, **41**, 3471–3481. doi: 10.1093/nar/gkt052

95 Bujara, M., Schumperli, M., Pellaux, R., Heinemann, M. *et al.* (2011) Optimization of a blueprint for in vitro glycolysis by metabolic real-time analysis. *Nat. Chem. Biol.*, **7**. doi: 10.1038/nchembio.54110.1038/NCHEMBIO.541

96 Lentini, R., Forlin, M., Martini, L., Del Bianco, C. *et al.* (2013) Fluorescent proteins and in vitro genetic organization for cell-free synthetic biology. *ACS Synth. Biol.*, **2**, 482–489. doi: 10.1021/sb400003y

97 Niederholtmeyer, H., Stepanova, V., and Maerkl, S.J. (2013) Implementation of cell-free biological networks at steady state. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 15985–15990. doi: 10.1073/pnas.1311166110

98 Hockenberry, A.J. and Jewett, M.C. (2012) Synthetic in vitro circuits. *Curr. Opin. Chem. Biol.*, **16**, 253–259. doi: 10.1016/j.cbpa.2012.05.179

99 Jewett, M.C., Fritz, B.R., Timmerman, L.E., and Church, G.M. (2013) In vitro integration of ribosomal RNA synthesis, ribosome assembly, and translation. *Mol. Syst. Biol.*, **9**, 678. doi: 10.1038/msb.2013.31

100 Nourian, Z. and Danelon, C. (2013) Linking genotype and phenotype in protein synthesizing liposomes with external supply of resources. *ACS Synth. Biol.*, **2**, 186–193. doi: 10.1021/sb300125z

101 Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74. doi: 10.1126/science.1093857

102 Shin, J., Jardine, P., and Noireaux, V. (2012) Genome replication, synthesis, and assembly of the bacteriophage T7 in a single cell-free reaction. *ACS Synth. Biol.*, **1**, 408–413. doi: 10.1021/sb300049p

103 Daube, S.S., Arad, T., and Bar-Ziv, R. (2007) Cell-free co-synthesis of protein nanoassemblies: tubes, rings, and doughnuts. *Nano Lett.*, **7**, 638–641.

16

Applying Advanced DNA Assembly Methods to Generate Pathway Libraries

Dawn T. Eriksen[1], *Ran Chao*[1], *and Huimin Zhao*[1,2]

[1] *University of Illinois at Urbana-Champaign, Department of Chemical and Biomolecular Engineering, 600 South Mathews Avenue, Urbana, IL 61801, USA*
[2] *University of Illinois at Urbana-Champaign, Departments of Chemistry, Biochemistry, and Bioengineering, 600 South Mathews Avenue, Urbana, IL 61801, USA*

16.1 Introduction

Pathways, which are cascades of biochemical reactions catalyzed by enzymes, maintain the vitality of all living organisms. These biochemical routes have been exploited to produce numerous commodities since early civilization, such as beer, wine, and cheese. With the advance of biotechnology, various genetic tools have become available for construction and manipulation of pathways to efficiently convert renewable feedstock to value-added compounds such as specialty chemicals, pharmaceuticals, and biofuels [1]. Microbial production of these compounds is usually enabled by overexpressing endogenous or heterologous enzymes of the corresponding pathways. However, overexpression of pathway enzymes alone can be insufficient for optimal metabolite production due to an imbalanced flux through the pathway [1, 2]. A typical symptom of flux imbalance is the accumulation of unwanted and even toxic intermediates [3, 4], which can be detrimental to the productivity of desired compounds. There is seldom a straightforward strategy to resolve the non-product accumulation because enzymes within the pathway are not independent; instead the enzymes are intertwined and cross-regulated among the pathway enzymes and among the cell's intricate metabolic networks. Due to this complexity, rationally engineering a pathway to improve its efficiency is a significant challenge. To this end, random approaches can be preferred over rational design in pathway engineering [5]. Random engineering approaches to optimize pathways generally screen through large and/or combinatorial pathway libraries. Pathway libraries have been constructed for diverse gene expression based on promoters of different strengths [6], varied intergenic regions affecting mRNA stability [4], or engineered ribosomal binding sites (RBSs) of diversified translational initiation rates [7].

In previous studies [4, 6–8], the pathway libraries were assembled by restriction digestion/ligation or overlap extension polymerase chain reaction (PCR).

Synthetic Biology: Parts, Devices and Applications, First Edition. Edited by Christina Smolke.

These traditional assembly methods were limited in complexity of design, being forced to rely on the multiple-cloning site (MCS) for pathway assembly, and had low assembly efficiency. In recent years, a number of new DNA assembly methods have been developed, such as DNA assembler [9], sequence and ligation-independent cloning (SLIC) [10], Gibson assembly [11], circular polymerase extension cloning (CPEC) [12], Golden Gate cloning [13], and BioBrick standards [14]. These advanced DNA assembly methods have ameliorated the design constraints on heterologous pathway construction and simplified the assembly of multi-gene metabolic pathways. The improved efficiency of these methods allows for larger and unbiased library creation, while the modularity of the methods greatly facilitates the generation of complex combinatorial libraries (Figure 16.1). The following chapter will include a brief description of the advanced assembly methods that could be applied to combinatorial pathway libraries. Some of the most recent work in pathway library generation using these methods will then be discussed as well.

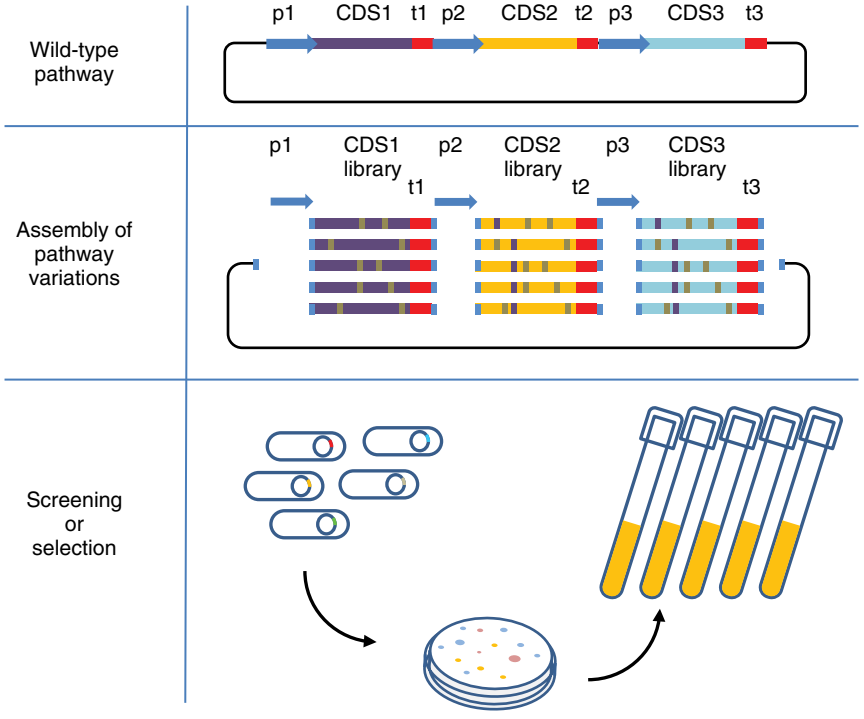


Figure 16.1 Overview of the combinatorial library approach for pathway improvement. When improving a multi-gene pathway, variations of the pathway components including promoters, RBSs, coding DNA sequences (CDSs), or transgenic regions are generated by either mutagenesis, homolog cloning, or *in silico* design (promoters and CDSs are used as examples in the figure). The diversified components are then assembled by various DNA assembly techniques to form a library of combinations. Cells hosting this pathway library will then be screened for the optima of the desired phenotype. Labels "p1-3" standard for promoters. Labels "t1-3" standard for terminators.

16.2 Advanced DNA Assembly Methods

The following methods have the potential to be used for pathway library creation (Table 16.1). The advanced assembly methods exploit diverse strategies for pathway construction such as homologous recombination, DNA polymerase extension, and advanced applications of restriction digestion/ligation.

The following assembly strategies are based on homologous recombination and DNA repair mechanisms: DNA assembler, Gibson assembly, and SLIC. In the DNA assembler strategy, the endogenous *in vivo* homologous recombination mechanism in yeast is used to create large pathways in a simple, one-step manner [9, 15, 16]. The DNA fragments to be assembled are PCR amplified by oligos designed with an 80-bp homologous region to the 5′ and 3′ neighboring DNA sequences within the pathway. The linear DNA fragments are co-transformed with the linear plasmid backbone into *Saccharomyces cerevisiae*, and the homologous regions are recognized by the endogenous homologous recombination machinery and "repaired" into a single DNA molecule.

Mimicking the *in vivo* homologous recombination mechanisms, *in vitro* assembly has been accomplished by SLIC and Gibson assembly. SLIC is a two-step DNA assembly method [10], which utilizes a 30-bp homology region. The linearized host vector and the insert DNA fragment are separately treated with T4 DNA polymerase in the absence of deoxynucleotide triphosphates (dNTPs), which chews back the 3′ terminal end. This generates a 5′ overhang that is homologous to the vector/insert. The second step involves addition of RecA and adenosine triphosphate (ATP), which can recombine the DNA fragments together into a single plasmid: any nicks generated are fixed after transformation. The Gibson assembly method [11] exploits a specific exonuclease to chew back the 5′ end to generate single-stranded complementary overhangs and ligases that are incorporated in the reaction mix to seal the DNA nicks. The DNA fragments are PCR amplified with 15–30 bp of homologous DNA regions to the 5′ and 3′ adjacent DNA sequences. In a single reaction, both vector and insert are subjected to T5 exonuclease that chews back the 5' ends of the DNA fragments, and then the polymerase and ligase combine the homologous ends of fragments to a single circular DNA molecule.

Table 16.1 Summary of different advanced DNA methods that could be used for combinatorial library generation.

Method	Type of reaction
SLIC	Exonuclease-based overhang generation and *in vivo* ligation
BioBrick standards	Step-wise modular restriction digestion and *in vitro* ligation
Golden Gate	Type IIs restriction enzyme digestion and *in vitro* ligation
DNA assembler	*In vivo* homologous recombination
Gibson assembly	Exonuclease-based overhang generation and *in vitro* ligation
CPEC	Overlap extension PCR

Homologous recombination is successful in DNA assembly, but basic polymerase extension mechanisms have also shown to be successful in the CPEC method to assemble DNA fragments into a plasmid [12]. The insert and vector are fused in an overlap extension PCR and circularize with extended overlapping stands, leaving only a nick in each strand. Then *Escherichia coli* repairs the nicks *in vivo* when transformed.

Another family of advanced DNA assembly techniques has been developed via the implementation of the type IIS endonucleases such as *Bsa*I, which cleave the DNA outside of their recognition sites, resulting in 5′ or 3′ DNA overhangs of nearly any user-defined nucleotide sequence [13, 17]. This strategy is more advanced than traditional restriction digestion/ligation method because it allows more flexibility in insertion location than cloning into the MCS on a plasmid. Use of type IIS endonucleases through the Golden Gate assembly method is a one-step reaction, which combines restriction digestion and ligation. This method has a high fragment assembly efficiency and proven to be effective in creating gene libraries [17]. A continuing area of research with this technique is investigating a more modular approach for pathway and pathway library construction [18, 19].

The need for modularity in gene and pathway cloning is becoming more significant with recent focuses on high-throughput DNA assembly and automation. One of the most established strategies for assembly standardization is the BioBrick system [14, 20–24]. The BioBrick and BglBrick standards (such as vectors, promoters, and RBS) rely on isocaudomer pairs of restriction enzymes to generate compatible cohesive ends and, upon ligation, result in a scar sequence that cannot be cleaved by either of the original restriction digests. DNA fragments flanked with these recognition sequences can be used for modular assembly of a pathway by iterative digestions and ligations.

Consideration of which assembly strategy to use for the generation of pathway libraries will greatly depend on the chassis, number of DNA fragments, and required assembly efficiency. *In vivo* homologous recombination is especially useful if the pathway is being expressed in *S. cerevisiae*. However, Gibson assembly and BioBrick standards are very useful if working in *E. coli*. Many DNA fragments to be assembled in the library can greatly decrease the assembly efficiency, which should be considered if a complex pathway is being investigated. If assembly efficiency is limiting, a strategy that allows for longer homology or linker region can be applied. Though no studies have linked library size to assembly strategies, some of the previous strategies might limit the library size, which could reduce the potential search space. Biases in assembly toward a certain gene or promoter can also reduce the potential search space. It is important to ensure that the library is diverse and random clones exhibit all potential genotypes of the library. One-pot assembly is also an important consideration, as iterative assemblies can be time consuming and can also reduce the potential library size.

16.3 Generation of Pathway Libraries

Combinatorial pathway library screening strategies, as compared with traditional pathway engineering strategies, can be more efficient in the identification of an optimized pathway. Traditional strategies optimize individual components

of the pathway one at a time to increase flux to the desired product [25–27], but pathway library screening strategies can tune multiple components of the pathway simultaneously. By varying multiple constituents concurrently, the likelihood of obtaining an optimized flux via balanced gene expression and protein activity within the pathway is increased. A more comprehensive exploration of the potential diversity of a target pathway can be achieved, which could identify unexpected synergistic effects [28, 29]. Many pathway optimization strategies are based on gene expression by varying promoter strength or RBS engineering. It is also possible to balance the flux through the pathway by exploring various combinations of enzymatic properties such as catalytic efficiency, cofactor specificity, stability, and substrate specificity. Currently, there are several examples of pathway libraries constructed through different advanced DNA assembly methods.

16.3.1 *In vitro* Assembly Methods

The Gibson assembly method was applied to generate a large combinatorial library of promoters and enzymes. The proof-of-concept pathway was the heterologous acetate utilization pathway in *E. coli*, comprised of an acetate kinase (*ackA*) and a phosphotransacetylase (pta) [30]. This combinatorial library was based on three promoter sequences with assorted strengths and four orthologous variants of both genes, generating 144 possible unique combinations of the promoters and genes. Each gene cassette was synthesized with an RBS, a terminator, and the promoter/gene variant. A unique 40-bp DNA linker sequence contains homologous DNA directly upstream and downstream of the gene at the terminal ends of the cassette (Figure 16.2). This linker region was used to ensure proper pathway sequence during assembly.

The total library size was approximately 10^4, affording 70-fold coverage of the 144 possible combinations. Investigation of the assembly efficiency showed that over 80% (30/37) of the selected clones harbored a correctly assembled pathway. Further sequencing analyses showed that of the thirty correctly assembled pathways, 60% (18/30) had recognizable promoter sequences. Of the possible 144 promoter/gene combinations, 14 unique combinations were present in the 18 positively identified pathways. A bias was noted toward a specific combination of genes from certain organisms, even though each gene fragment was assembled in equal combinations. This bias could have been the result of an assembly bias, or it could be the result of a screening bias, as the library was screened on acetate and these genes could be the most efficient for acetate utilization in *E. coli.*

The Gibson assembly was also used by Coussement and coworkers in another example of creating a combinatorial library of transcription, translation, and protein sequence variability [31]. This strategy utilized a single-stranded assembly to introduce diversity in the double-stranded DNA of the promoter, RBS, and/or coding sequences. Optimization of the assembly found that two oligonucleotide fragments of similar lengths provided a nearly 100% efficiency of assembly. More DNA fragments or fragments of different lengths lowered the assembly efficiency. Promoter, RBS, and protein libraries using a single gene were all proven to have a large linear range and had diverse expression and activity. The assembly was tested for combinatorial pathway libraries using the

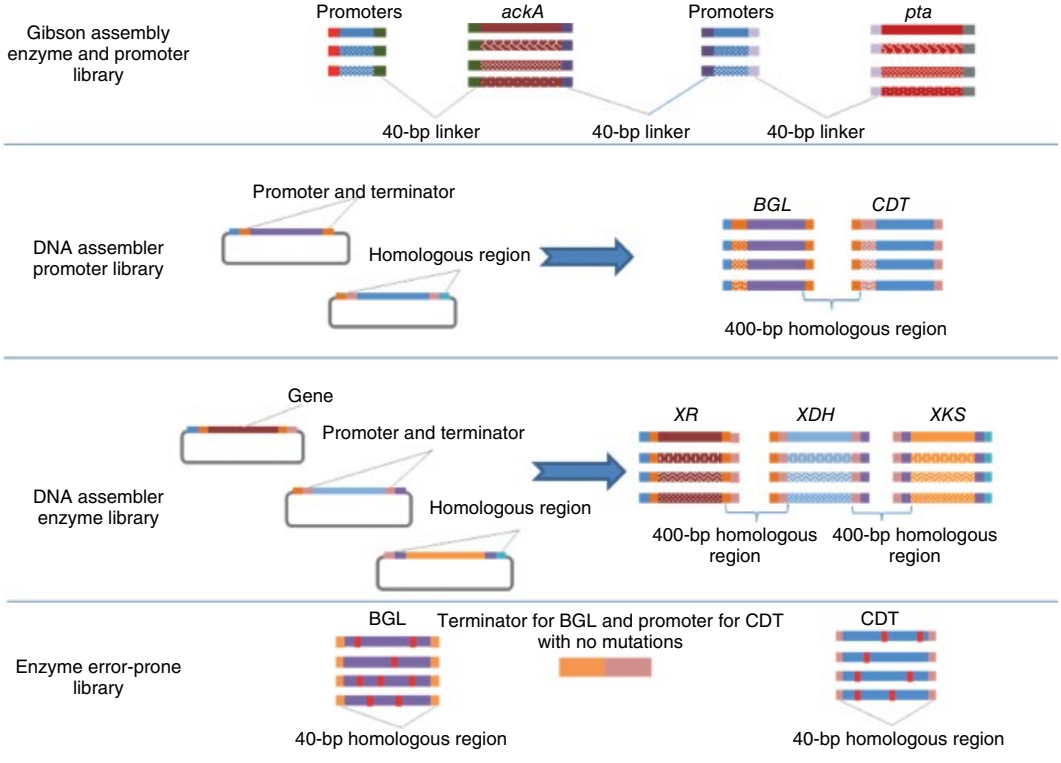


Figure 16.2 Preparation of DNA fragments for large library generation. Each unique design represents a unique promoter or gene. Varied strength promoters, orthologous genes, or mutated pathway components generate diversity. If the DNA is assembled with homology regions, upstream and downstream of the DNA fragment of interest, the pathway can assemble properly into many different combinations. Each strategy has incorporated different lengths of homology, which can contribute to the efficiency of correct assembly. These DNA fragments are then subjected to the desired DNA assembly reaction with the linearized vector and transformed into the host.

reporter genes mKate2 and sfGfp. The promoter library was tested using this fluorescent pathway. One hundred and eighty-eight clones were randomly picked and profiled for complete representation of the potential expression landscape. Theoretical library size was 4^{32} and these 188 clones revealed a good representation of diversity. This pathway optimization is based on short fragment assembly of 50–150-bp assembly and has not been applied to larger DNA fragments. This strategy is efficient for shorter promoter regions of *E. coli* and point mutations of targeted protein engineering, which is one of the preferred strategies for protein engineering. However, pathways incorporating diversity in larger DNA fragments such as yeast promoters and other protein engineering strategies could not be accomplished through this current method.

The Gibson assembly was also utilized by Lee *et al.* to optimize a multienzyme pathway in the absence of a high-throughput assay [32]. This study took the pathway libraries assembly one step further and incorporated computational modeling to reduce the large search library that must be screened. For assembly, standardized vectors were constructed based on principles of the BglBrick-style cloning of protein fusions. The expression cassettes were flanked by pairs of homology sequences (20 bp) derived from yeast barcodes to allow for correct sequence assembly. Each promoter used was proven to work independently of DNA sequence directly downstream of it. Three-gene library assemblies resulted in 25–33% miss-assembly. Library assembly was tested in a three-gene fluorescent protein library, with a theoretical library size of 125. The triple library was shown to cover the complete three-dimensional expression space.

To apply this assembly to a pathway and construct a predictive model, the five-gene violacein biosynthetic pathway was utilized, resulting in a theoretical combinatorial library size of 3125. Ninety-one random transformants from the colony were characterized for geno- and phenotypic data. A linear regression model was then constructed from this data and used to predict optimal phenotypes. The authors suggest that a low sampling rate of 1–2% of the library could be sufficient for generating a predictive model. Four models were constructed for different intermediates and branched products of the violacein pathway. The model predictions and empirical data were high, with Pearson correlation coefficients being between 0.77 and 0.92 for the specific targets. The model was used to predict the top five expression-level combinations. These combinations were individually cloned and tested to determine if the desired product had increased production with the predicted expression levels. The model was able to predict and identify the expression level to yield the desired product with the highest production from the pathway.

A BioBrick-like assembly strategy was used in a combinatorial library of engineered RBSs [33]. This iterative assembly process utilizes the chloramphenicol resistance cassette paired with the library of RBS sequences. The resistance cassette is flanked by restriction digests and then can easily be removed to incorporate the next target gene and RBS library. To determine if the strategy could yield a library that spanned a multidimensional expression space, three reporter genes were used in a synthetic operon: CFP, YP, and mCherry. The RBS modulation was shown to span 100-fold in each dimension of the expression space. The seven-gene carotenoid biosynthesis pathway with the end product of astaxanthin

was used as a proof-of-concept study for this assembly in pathways. The theoretical library was 6^7 possible RBS combinations, and nearly 25 000 clones were visually screened, which is only 10% of the potential library. Through visual screening of the colonies' color of astaxanthin, 500 colonies were picked for further analysis. Fifty clones were identified to have the most intense color and screened for highest astaxanthin production through high-performance liquid chromatography (HPLC). This strategy yielded a clone with fourfold higher astaxanthin production than the wild-type pathway.

The aforementioned studies have all involved random, large pathway libraries. A new BioBrick standard platform, the ePathBrick system, allows for assembly of specific pathways, with the ability to vary specific components [34]. The ePathBrick system is a pathway fine-tuning toolkit that consists of a number of BioBrick-compatible plasmids with characterized regulatory signal elements. With this system, Xu and coworkers demonstrated a modular engineering approach for significant titer improvement of a multi-gene fatty acid metabolic pathway by fine-tuning gene expression through plasmid copy number and RBS engineering [35]. The *E. coli* fatty acid biosynthetic pathway was apportioned and overexpressed in three separate modules. These modules were successfully expressed on compatible ePathBrick vectors with varying plasmid copy numbers. The total fatty acid production was optimized by overexpressing each module on high, medium, or low copy number plasmids. Nine independent pathways were constructed through the ePathBrick standards and analyzed for fatty acid production. As has been noted before in product titer, the highest gene expression is not always optimal [6]. The greatest increase in fatty acid production occurred only when the final module was expressed highly, combined with a lower expression in the other modules. The balanced gene expression pathway produced a fourfold increase in fatty acid titer compared to the lowest-producing pathway. Similarly, three different strength RBSs were also tested in the modules, and a balance between strong and medium strength RBSs improved fatty acid production by twofold. This type of strategy can illuminate bottlenecks in the pathway. This study exemplified the importance of high concentrations of malonyl-CoA in fatty acid production.

A randomized BioBrick strategy has also been developed, which combines the power of Gibson assembly and the modularity of the BioBrick standards [36]. In this method, all promoters, RBSs, and transcriptional terminators were randomized within the pathway. These modular DNA fragments were derived from PCR-amplified BioBricks, and each component was cloned with 18–28-bp linkers of homologous DNA regions to the 5′ and 3′ DNA. Three promoters, three RBSs, and three terminators were simultaneously randomized for the three-gene pathway for the lycopene biosynthetic pathway, generating a library of nearly 20 000 unique clones. The library was assembled through Gibson assembly and was screened on plates for the orange-colored lycopene product. Of the red–orange colored colonies, 12 were selected, and DNA sequencing analysis demonstrated that 7/8 randomized pathways were distinct and four pathways had deletions. The study cautions the metabolic burden placed on the cells during the library screening that could have caused the mutations.

16.3.2 *In vivo* Assembly Methods

E. coli does not have robust and efficient homologous recombination machinery; therefore *in vitro* assembly methods are highly needed. In contrast, plants and yeast have very vigorous and efficient homologous recombination machinery, allowing for facile pathway library creation *in vivo*. Two divergent strategies for *in vivo* homologous recombination have been developed: chromosomal integration and plasmid assembly.

16.3.2.1 *In vivo* Chromosomal Integration

Wingler and Cornish established a reiterative recombination method for the *in vivo* assembly of multi-gene pathway libraries directly into the chromosome [37]. The strategy utilized a pair of alternating orthogonal endonucleases and selectable markers. Homologous recombination and gap repair were used to construct a plasmid containing the gene of interest, marker, and endonuclease, which were recombined into an acceptor strain. This acceptor strain carries a predefined target locus for integration into the chromosome. Galactose-induced expression of the endonuclease cleaves the double-stranded DNA, triggering the homologous recombination and leading to integration of the gene of interest and the auxotrophic marker into the chromosome. The strains are then selected for the new auxotrophic marker and cured against excess donor plasmid. The proof of concept for pathway integration and mock library assembly was demonstrated using the lycopene biosynthetic pathway (*crtE*, *crtB*, and *crtI*). A large library of over 10^4 was assembled: the mock library contained various ratios of *crtB* and *crtI* alleles that contained either nonsense or silent mutations, which would produce working or interrupted pathways. The diversity could be judged based on the actual and theoretical percentages of working pathways versus interrupted pathways, visualized based on the color of the colonies on the plate. Each library had the expected percentage of working pathways, indicating a non-biased library assembly into the chromosome.

Pathway library strategies have also been established in plant biotechnology to study secondary metabolites [38, 39]. Engineering secondary metabolism in plants can be a daunting task considering the complexity of the target pathways, which could have multiple branches, multifunctional and/or compartmentalized enzymes, and complex feedback inhibition. Zhu *et al.* established a novel method for the combinatorial nuclear transformation of multiple genes into a plant, generating a pathway library to simplify the study of multiple variables of secondary metabolites [38, 39]. Carotenoid production in cereal grains was used as a proof of concept. Embryos of the cereal-grain white maize were bombarded with metal particles coated with six unique constructs, consisting of a selection marker and five carotenogenic genes. The resultant library consisted of any combination of one or more expression phenotype from any of the five genes. This method of multiple gene transformation and pathway library screening allowed the identification of rate-limiting steps in the carotenogenic pathway. Total carotenoid production in cereal grains was improved 140-fold based on a unique combination identified from this multi-gene pathway library strategy.

16.3.2.2 *In vivo* Plasmid Assembly and One-Step Optimization Libraries

Chromosomal integration has been successful in pathway library creation, but assembling the pathways into a plasmid is also advantageous. A plasmid is a DNA molecule that can be easily transported across strains, which is an important characteristic to consider when excluding the possibility that the observed improvements are not a result of off-target genome modification.

An example of plasmid-based pathway libraries was constructed by the DNA assembly method and focused on a combinatorial library of different promoter strengths for all the genes within the library [40]. As a proof of concept in pathway library generation, the xylose and cellobiose utilization pathways for ethanol production were optimized. Efficient utilization of these biomass sugars is critical for economically feasible biofuel production. Promoters *PDC1*, *ENO2*, and *TEF1* were mutagenized through nucleotide analog-based error-prone PCR to induce a very high mutation rate and produce promoters of various strengths. After mutagenesis, mutants for each promoter were assayed through fluorescence protein expression, and 10 promoters of defined strengths were selected for library construction. These 10 promoters in each position of the library resulted in a theoretical library size of 10^2 and 10^3 for the cellobiose and xylose utilization libraries, respectively. Each mutant promoter was cloned into a helper plasmid that contained 400-bp sequences homologous to the 5′ DNA region (Figure 16.2). The mutant promoter/gene expression cassettes were co-transformed into a yeast strain with a total library size of 10^5. To confirm the diversity of the library, over 40 individual colonies from each library were screened from an antibiotic selection marker for plasmid-pathway assembly and not based on sugar utilization. Each colony from this plasmid marker selection exhibited a unique growth curve on its respective carbon source, which was indicative of a diverse library.

Improved sugar utilization was visualized in a high-throughput manner by inspection of colony size on agar plates, wherein larger colony sizes were suggestive of faster sugar utilization and improved growth. In the xylose utilization pathway, a very efficient mutant pathway was identified in a single step. This pathway conferred a xylose consumption rate of $0.73\,g\,l^{-1}\,h^{-1}$, comparable with some of the fastest xylose consumption rates from strains that had been subjected to multiple generations of optimization strategies. The strain harboring the wild-type pathway did not produce any ethanol, while the mutant pathway conferred an ethanol productivity of $0.17\,g\,l^{-1}\,h^{-1}$. In the cellobiose utilization strategy, the strain harboring the optimized pathway yielded a 5.4-fold improved cellobiose utilization rate and a 5.3-fold increase in ethanol productivity.

A similar pathway library strategy created a combinatorial library of homologous enzymes of the xylose utilization pathway, with fix-strength promoters [41]. The fungal xylose utilization pathway has been shown to be especially sensitive to cofactor imbalances and unbalanced enzyme expression [42–45]. A total theoretical library size of 8360 possible unique combinations of homologous enzymes for each of the five genes in the pathway was constructed through homologous recombination. Each enzyme was characterized to show varied activities and cofactor dependencies. The gene sequences were cloned into helper plasmid expression cassettes, containing promoters, terminators, and at

least a 400-bp region homologous to the 5′ and 3′ DNA regions of the pathway at the termini of the expression cassette (Figure 16.2). The expression cassettes were transformed into the three different yeast strains with an average library size of 1.3×10^4. To confirm library diversity and screening for optimal pathways, the same strategies established in the promoter-based library were applied [40]. Sequencing results of random colonies showed that all the genes were recognizable with no major mutations or hybrids, resulting in a 100% efficiency, and there was no significant bias toward a certain gene. The same library was screened in three different strains, and a unique combination of genes was discovered to be optimal in each individual strain. This unique combination for each strain is attributed to the different metabolic background of the strains and availability of precursors or cofactors.

16.3.2.3 *In vivo* Plasmid Assembly and Iterative Multi-step Optimization Libraries

Directed evolution, an iterative multistep optimization strategy, is an established strategy that is a very powerful technique in synthetic biology for optimizing protein activity [5, 46]. Application of the strategy has been expanded to include pathway-scale transcriptional engineering and protein engineering through the following pathway library studies. The directed evolution strategy on the pathway scale is particularly powerful because it allows for the optimal flux to be identified with no a priori information about pathway bottlenecks or specifics about the pathway enzymes. This directed evolution strategy on the pathway scale allows for all components to be screened/selected for a balanced activity, not just for high activity.

Yuan and coworkers applied directed evolution to mutant promoter pathway libraries of the cellobiose utilization [47]. An average mutation rate of 12–16-nucleotide substitutions per kilobase for each mutagenized promoter was obtained. The pathway genes were not mutagenized, and these non-mutated DNA fragments were co-transformed with the error-prone promoter library and a linearized vector for a total library size of 10^4. The pathway phenotype improvement was assessed by fast sugar utilization, visualized by large colonies on agar plates. The first round of directed evolution identified a strain with a 5.7-fold increase in cellobiose consumption rate and a 5.5-fold increase in ethanol productivity. The further rounds of evolution yielded incremental subsequent increases (Figure 16.3). After characterizing the mutant promoters, it was found that the expression level ratios had significantly changed. While the parent BGL:CDT (β-glucosidase/cellodextrin transporter) relative expression ratio was 13.8 : 1, the first round of mutagenesis altered the ratio to 2.5 : 1. This significant increase in relative CDT expression suggested that this protein expression was a bottleneck.

Pathway-scale protein engineering strategies were also applied using homologous recombination [48]. In this study, both the BGL and CDT proteins were coevolved for balanced activity in a directed evolution manner. One amino acid substitution per protein was introduced through error-prone PCR, yielding a theoretical total library size of 9.9×10^6. No gene expression elements were mutagenized in this strategy and therefore were PCR amplified into the pathway

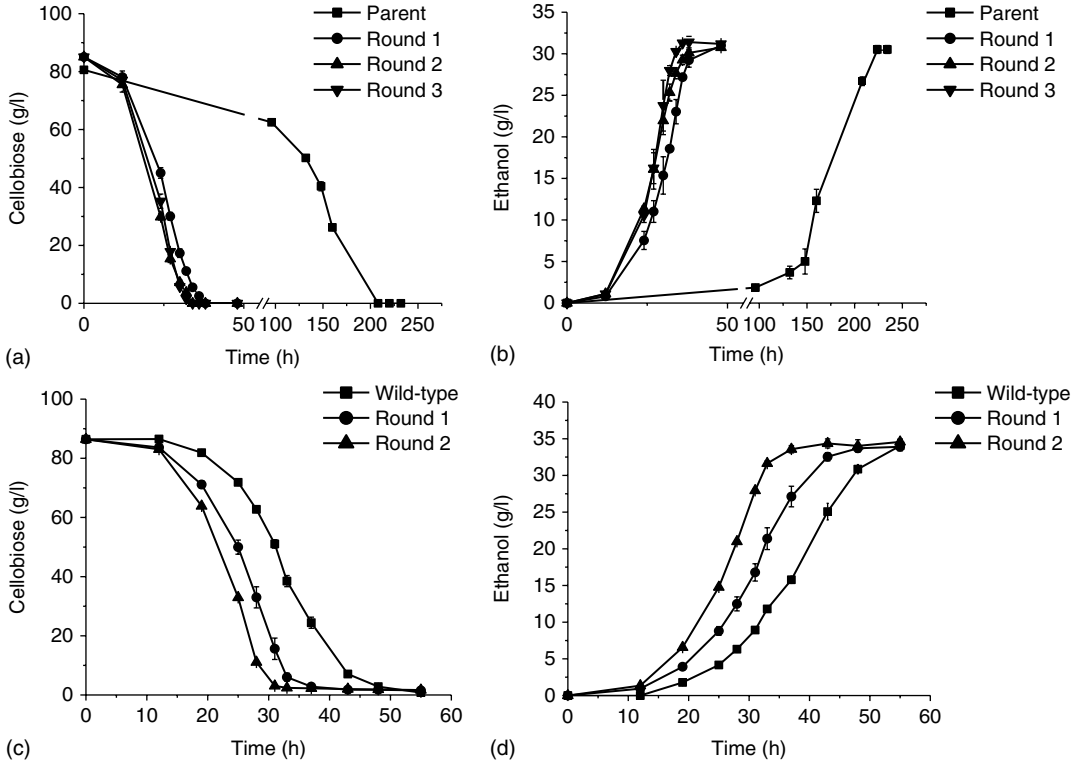


Figure 16.3 Fermentation profiles of the evolutionary rounds for the pathway libraries. (a,b) Cellobiose consumption and ethanol production of the cellobiose utilization pathway from the promoter-based directed evolution. The black square represents the parent pathway with no mutations in the *PDC1* and *ENO2* promoter. The circles are the first round of error-prone PCR of both promoters. The triangles represent the second and final rounds of directed evolution mutagenesis. (c,d) Cellobiose consumption and ethanol production of the cellobiose utilization pathway from the protein-based directed evolution. The black square represents the wild-type pathway with no mutations in the β-glucosidase and the cellodextrin transporter. The circle is the first round of error-prone PCR of both proteins. The triangle represents the second round of directed evolution.

separately (Figure 16.2). The total library size screened was 10^4 and was screened for strains harboring a pathway that conferred a fast growth on cellobiose, visualized through large colonies on cellobiose agar plates. In this study, two rounds of directed evolution identified a mutant pathway that conferred a 47% increase in growth rate on cellobiose and a 64% increase in ethanol productivity (Figure 16.3). As all proteins of the pathway were coevolved, mutations were found in each protein from every round and characterized to understand why the pathway conferred an improved phenotype. The BGL mutants were shown to have improved cellobiose specificity and activity. The CDT mutants had an overall higher activity, associated with a higher V_{max}.

16.4 Conclusions and Prospects

Advanced DNA assembly methods have allowed scientists and engineers extraordinary freedom in constructing pathways, greatly facilitating advances in pathway library generation. Pathway optimization through whole pathway libraries has expanded the potential diversity and possibilities for improving pathway phenotype. Furthermore, high efficiency and modularity of these advanced DNA assembly methods make *in silico* design [49] and automated assembly [50] of these libraries possible. Large combinations of library components can be individually constructed by robotic platforms and investigated by high-throughput screening for extensive investigations of improved pathway phenotypes. Despite the rapid progress of DNA assembly technologies, widespread application of pathway libraries is currently limited by high-throughput screening. Without the ability to easily and economically quantify the phenotype of interest, these large-scale pathway libraries will not be able to fulfill their maximum potential. Future high-throughput screening methods could be realized through microfluidic devices, with the ability to screen up to 10^8 clones per day [51, 52]. Biosensors also have potential in high-throughput screening, as shown by a number of transcription factor-based biosensors that have been engineered to detect small molecules. These biosensors can link the small molecule concentration to an easily measurable signal such as fluorescence and cell growth via gene circuits [53–56]. Though there are challenges, the potential of using advanced DNA assembly methods to create pathway libraries to significantly improve microbial cell production of fuels and chemicals is significant, and future pathway engineering methods will benefit from these strategies.

Definitions

Pathway Coordinated heterologous and/or endogenous enzymatic reactions

Pathway engineering A research area that specializes in modifying or optimizing components of an enzymatic pathway for improved phenotype

Pathway optimization Strategies to improve the overall performance of an enzymatic pathway

Pathway libraries A collection of mutant enzymatic pathways wherein multiple components (RBS, promoters, enzymes) within the pathway have simultaneously been mutated

Directed evolution An evolutionary process for engineering biological systems that mimics Darwinian evolution *in vitro and in vivo*: rounds of random mutations are incorporated into the DNA sequence and selected for improved phenotype in an iterative fashion

DNA assembly The process to conjoin several DNA fragments to create a large DNA molecule

References

1 Du, J., Shao, Z., and Zhao, H. (2011) Engineering microbial factories for synthesis of value-added products. *J. Ind. Microbiol. Biotechnol.*, **38**, 873–890.

2 Jones, K.L., Kim, S.W., and Keasling, J.D. (2000) Low-copy plasmids can perform as well as or better than high-copy plasmids for metabolic engineering of bacteria. *Metab. Eng.*, **2**, 328–338.

3 Rohlin, L., Oh, M.K., and Liao, J.C. (2001) Microbial pathway engineering for industrial processes: evolution, combinatorial biosynthesis and rational design. *Curr. Opin. Microbiol.*, **4**, 330–335.

4 Pfleger, B.F., Pitera, D.J., Smolke, C.D., and Keasling, J.D. (2006) Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. *Nat. Biotechnol.*, **24**, 1027–1032.

5 Cobb, R.E., Chao, R., and Zhao, H. (2013) Directed evolution: past, present and future. *AIChE J.*, **59**, 1432–1440.

6 Alper, H., Fischer, C., Nevoigt, E., and Stephanopoulos, G. (2005) Tuning genetic control through promoter engineering. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 12678–12683.

7 Salis, H.M., Mirsky, E.A., and Voigt, C.A. (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.*, **27**, 946–950.

8 Wang, C., Oh, M.K., and Liao, J.C. (2000) Directed evolution of metabolically engineered *Escherichia coli* for carotenoid production. *Biotechnol. Prog.*, **16**, 922–926.

9 Shao, Z.Y., Zhao, H., and Zhao, H.M. (2009) DNA assembler, an *in vivo* genetic method for rapid construction of biochemical pathways. *Nucleic Acids Res.*, **37**, e16.

10 Li, M.Z. and Elledge, S.J. (2007) Harnessing homologous recombination *in vitro* to generate recombinant DNA via SLIC. *Nat. Methods*, **4**, 251–256.

11 Gibson, D.G., Young, L., Chuang, R.Y., Venter, J.C., Hutchison, C.A. *et al.* (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, **6**, 343–347.

12 Quan, J. and Tian, J. (2009) Circular polymerase extension cloning of complex gene libraries and pathways. *PLoS One*, **4**, e6441.

13 Engler, C., Kandzia, R., and Marillonnet, S. (2008) A one pot, one step, precision cloning method with high throughput capability. *PLoS One*, **3**, e3647.

14 Shetty, R.P., Endy, D., and Knight, T.F. Jr. (2008) Engineering BioBrick vectors from BioBrick parts. *J. Biol. Eng.*, **2**, 1–12.

15 Shao, Z., Luo, Y., and Zhao, H. (2011) Rapid characterization and engineering of natural product biosynthetic pathways via DNA assembler. *Mol. Biosyst.*, **7**, 1056–1059.

16 Kuijpers, N.G., Solis-Escalante, D., Bosman, L., van den Broek, M., Pronk, J.T. *et al.* (2013) A versatile, efficient strategy for assembly of multi-fragment expression vectors in *Saccharomyces cerevisiae* using 60 bp synthetic recombination sequences. *Microb. Cell Fact.*, **12**, 47.

17 Engler, C., Gruetzner, R., Kandzia, R., and Marillonnet, S. (2009) Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PLoS One*, **4**, e5553.

18 Werner, S., Engler, C., Weber, E., Gruetzner, R., and Marillonnet, S. (2012) Fast track assembly of multigene constructs using Golden Gate cloning and the MoClo system. *Bioeng. Bugs*, **3**, 38–43.

19 Weber, E., Engler, C., Gruetzner, R., Werner, S., and Marillonnet, S. (2011) A modular cloning system for standardized assembly of multigene constructs. *PLoS One*, **6**, e16765.

20 Knight, T. (2003) Idempotent Vector Design for Standard Assembly of Biobricks, MIT Synthetic Working Group, http://web.mit.edu/synbio/release/docs/biobricks.pdf.

21 Canton, B., Labno, A., and Endy, D. (2008) Refinement and standardization of synthetic biological parts and devices. *Nat. Biotechnol.*, **26**, 787–793.

22 Vick, J.E., Johnson, E.T., Choudhary, S., Bloch, S.E., Lopez-Gallego, F. *et al.* (2011) Optimized compatible set of BioBrick vectors for metabolic pathway engineering. *Appl. Microbiol. Biotechnol.*, **92**, 1275–1286.

23 Lee, T.S., Krupa, R.A., Zhang, F., Hajimorad, M., Holtz, W.J. *et al.* (2011) BglBrick vectors and datasheets: a synthetic biology platform for gene expression. *J. Biol. Eng.*, **5**, 1–12.

24 Anderson, J.C., Dueber, J.E., Leguia, M., Wu, G.C., Goler, J.A. *et al.* (2010) BglBricks: a flexible standard for biological part assembly. *J. Biol. Eng.*, **4**, 1–12.

25 Leonard, E., Ajikumar, P.K., Thayer, K., Xiao, W.H., Mo, J.D. *et al.* (2010) Combining metabolic and protein engineering of a terpenoid biosynthetic pathway for overproduction and selectivity control. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 13654–13659.

26 Matsushika, A., Watanabe, S., Kodaki, T., Makino, K., Inoue, H. *et al.* (2008) Expression of protein engineered $NADP^+$-dependent xylitol dehydrogenase increases ethanol production from xylose in recombinant *Saccharomyces cerevisiae*. *Appl. Microbiol. Biotechnol.*, **81**, 243–255.

27 Nakamura, C.E. and Whited, G.M. (2003) Metabolic engineering for the microbial production of 1,3-propanediol. *Curr. Opin. Biotechnol.*, **14**, 454–459.

28 Wang, H.H., Isaacs, F.J., Carr, P.A., Sun, Z.Z., Xu, G. *et al.* (2009) Programming cells by multiplex genome engineering and accelerated evolution. *Nature*, **460**, 894–898.

29 Alper, H., Miyaoku, K., and Stephanopoulos, G. (2005) Construction of lycopene-overproducing *E. coli* strains by combining systematic and combinatorial gene knockout targets. *Nat. Biotechnol.*, **23**, 612–616.

30 Ramon, A. and Smith, H.O. (2011) Single-step linker-based combinatorial assembly of promoter and gene cassettes for pathway engineering. *Biotechnol. Lett*, **33**, 549–555.
31 Coussement, P., Maertens, J., Beauprez, J., Van Bellegem, W., and De Mey, M. (2014) One step DNA assembly for combinatorial metabolic engineering. *Metab. Eng.*, **23**, 70–77.
32 Lee, M.E., Aswani, A., Han, A.S., Tomlin, C.J., and Dueber, J.E. (2013) Expression-level optimization of a multi-enzyme pathway in the absence of a high-throughput assay. *Nucleic Acids Res.*, **41**, 10668–10678.
33 Zelcbuch, L., Antonovsky, N., Bar-Even, A., Levin-Karp, A., Barenholz, U. *et al.* (2013) Spanning high-dimensional expression space using ribosome-binding site combinatorics. *Nucleic Acids Res.*, **41**, e98.
34 Xu, P., Vansiri, A., Bhan, N., and Koffas, M.A. (2012) EPathBrick: a synthetic biology platform for engineering metabolic pathways in *E. Coli*. *ACS Synth. Biol.*, **1**, 256–266.
35 Xu, P., Gu, Q., Wang, W., Wong, L., Bower, A.G. *et al.* (2013) Modular optimization of multi-gene pathways for fatty acids production in *E. coli*. *Nat. Commun.*, **4**, 1409.
36 Sleight, S.C. and Sauro, H.M. (2013) Randomized BioBrick assembly: a novel DNA assembly method for randomizing and optimizing genetic circuits and metabolic pathways. *ACS Synth. Biol.*, **2**, 506–518.
37 Wingler, L.M. and Cornish, V.W. (2011) Reiterative recombination for the *in vivo* assembly of libraries of multigene pathways. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 15135–15140.
38 Zhu, C., Naqvi, S., Breitenbach, J., Sandmann, G., Christou, P. *et al.* (2008) Combinatorial genetic transformation generates a library of metabolic phenotypes for the carotenoid pathway in maize. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 18232–18237.
39 Farre, G., Naqvi, S., Sanahuja, G., Bai, C., Zorrilla-Lopez, U. *et al.* (2012) Combinatorial genetic transformation of cereals and the creation of metabolic libraries for the carotenoid pathway. *Methods Mol. Biol.*, **847**, 419–435.
40 Du, J., Yuan, Y., Si, T., Lian, J., and Zhao, H. (2012) Customized optimization of metabolic pathways by combinatorial transcriptional engineering. *Nucleic Acids Res.*, **40**, e142.
41 Kim, B., Du, J., Eriksen, D.T., and Zhao, H. (2013) Combinatorial design of a highly efficient xylose utilizing pathway for cellulosic biofuels production in *Saccharomyces cerevisiae*. *Appl. Environ. Microbiol.*, **79**, 931–941.
42 Jin, Y.S. and Jeffries, T.W. (2003) Changing flux of xylose metabolites by altering expression of xylose reductase and xylitol dehydrogenase in recombinant *Saccharomyces cerevisiae*. *Appl. Biochem. Biotechnol.*, **105–108**, 277–786.
43 Karhumaa, K., Fromanger, R., Hahn-Hagerdal, B., and Gorwa-Grauslund, M.F. (2007) High activity of xylose reductase and xylitol dehydrogenase improves xylose fermentation by recombinant *Saccharomyces cerevisiae*. *Appl. Microbiol. Biotechnol.*, **73**, 1039–1046.
44 Jin, Y.S., Ni, H., Laplaza, J.M., and Jeffries, T.W. (2003) Optimal growth and ethanol production from xylose by recombinant *Saccharomyces cerevisiae*

require moderate D-xylulokinase activity. *Appl. Environ. Microbiol.*, **69**, 495–503.

45 Matsushika, A., Watanabe, S., Kodaki, T., Makino, K., and Sawayama, S. (2008) Bioethanol production from xylose by recombinant *Saccharomyces cerevisiae* expressing xylose reductase, NADP(+)-dependent xylitol dehydrogenase, and xylulokinase. *J. Biosci. Bioeng.*, **105**, 296–299.

46 Cobb, R.E., Si, T., and Zhao, H. (2012) Directed evolution: an evolving and enabling synthetic biology tool. *Curr. Opin. Chem. Biol.*, **16**, 285–291.

47 Yuan, Y. and Zhao, H. (2013) Directed evolution of a highly efficient cellobiose utilizing pathway in an industrial *Saccharomyces cerevisiae* strain. *Biotechnol. Bioeng.*, **110**, 2874–2881.

48 Eriksen, D., Hsieh, H., Lynn, P., and Zhao, H. (2013) Directed evolution of a cellobiose utilization pathway in *Saccharomyces cerevisiae* by simultaneously engineering multiple proteins. *Microb. Cell Fact.*, **12**, 1–10.

49 Hillson, N.J., Rosengarten, R.D., and Keasling, J.D. (2012) j5 DNA assembly design automation software. *ACS Synth. Biol.*, **1**, 14–21.

50 Linshiz, G., Stawski, N., Poust, S., Bi, C., Keasling, J.D. *et al.* (2012) PaR-PaR laboratory automation platform. *ACS Synth. Biol.*, **2**, 216–222.

51 Uhlen, M. and Svahn, H.A. (2011) Lab on a chip technologies for bioenergy and biosustainability research. *Lab Chip*, **11**, 3389–3393.

52 Guo, M.T., Rotem, A., Heyman, J.A., and Weitz, D.A. (2012) Droplet microfluidics for high-throughput biological assays. *Lab Chip*, **12**, 2146–2155.

53 Dietrich, J.A., Shis, D.L., Alikhani, A., and Keasling, J.D. (2013) Transcription factor-based screens and synthetic selections for microbial small-molecule biosynthesis. *ACS Synth. Biol.*, **2**, 47–58.

54 Pfleger, B.F., Pitera, D.J., Newman, J.D., Martin, V.J., and Keasling, J.D. (2007) Microbial sensors for small molecules: development of a mevalonate biosensor. *Metab. Eng.*, **9**, 30–38.

55 Zhang, F., Carothers, J.M., and Keasling, J.D. (2012) Design of a dynamic sensor-regulator system for production of chemicals and fuels derived from fatty acids. *Nat. Biotechnol.*, **30**, 354–359.

56 Zhang, F. and Keasling, J.D. (2011) Biosensors and their applications in microbial metabolic engineering. *Trends Microbiol.*, **19**, 323–329.

17

Synthetic Biology in Immunotherapy and Stem Cell Therapy Engineering

Patrick Ho and Yvonne Y. Chen

University of California, Department of Chemical and Biomolecular Engineering, 420 Westwood Plaza, Boelter Hall 5532, Los Angeles, CA 90095, USA

17.1 The Need for a New Therapeutic Paradigm

The advent of the germ theory of disease in the late nineteenth century marked a watershed in the history of medicine and heralded the development of modern pharmaceuticals. The work of Louis Pasteur, Robert Koch, and fellow microbiologists elucidated the bacterial and viral origins of common and often fatal diseases such as cholera and puerperal fever and motivated the development of myriad small molecule-based pharmaceuticals and viral vaccines that specifically targeted infectious agents. As epidemics such as smallpox and polio came under control, new classes of diseases that do not have simple biological causes gradually took center stage. Chronic and complex illnesses such as diabetes, cardiovascular diseases, and cancers supplanted infectious diseases as the dominant scourges in developed countries after the Second World War. In response to this changing landscape of medical challenges, pioneers in molecular biology and genetic engineering launched a new paradigm of pharmaceutical development and, beginning in the 1980s, produced the first biologics: monoclonal antibodies such as trastuzumab (Herceptin) [5, 6] and recombinant protein therapeutics such as synthetic insulin and erythropoietin [7, 8]. Today, the twin pillars of small molecules and biologics continue to serve as the pharmaceutical arsenal of modern medicine, complemented by non-biochemical methods such as medical devices and surgical intervention.

Despite modern advancements in diagnostics and therapeutics, several debilitating diseases have remained essentially incurable. In particular, cancer has steadily risen through the ranks of fatal diseases over the past several decades, with prominent examples including pancreatic and small cell lung cancers, each with an overall relative 5-year survival rate of 7% [9]. Glioblastoma, the most common type of primary brain tumors, has a median survival period of less than 15 months [10, 11]. Unlike well-characterized infectious diseases and metabolic disorders such as diabetes, the conditions highlighted previously do not present simple biological causes or deficiencies that can be easily eliminated or compensated by chemical drugs and biologics. Cancer cells are characterized by genomic

Synthetic Biology: Parts, Devices and Applications, First Edition. Edited by Christina Smolke.

Table 17.1 Major categories of cell-based immunotherapies and application areas currently under investigation.

Cell type	Major application areas
Effector and memory T cells	Cancers, viral infections
Regulatory T cells	Autoimmune diseases, inflammatory diseases
Myeloid-derived suppressor cells	Autoimmune diseases, inflammatory diseases
Dendritic cells	Cancer vaccines
Natural killer cells	Cancers, viral infections

instabilities that enable them to escape individual therapeutic strategies through genetic hypermutation [12]. Furthermore, in cases such as glioblastoma, diseased cells can be situated in a protected niche (e.g., behind the blood-brain barrier (BBB)) that is both inaccessible to chemical and biological therapeutics and incompatible with complete surgical resection [13]. Finally, diseased cells often closely resemble healthy tissues on the surface and lack unique molecular markers that allow precise identification by drug molecules. As a result, strategies including chemotherapy and antibody therapeutics often lead to severe off-target or "on-target, off-tumor" toxicities [5, 14].

Complex, dynamic diseases call for a new category of therapeutics that can actively sense and process multiple input signals and respond to changing disease landscapes with multipronged therapeutic outputs [1–4]. Cellular therapies represent a new platform for the treatment of currently intractable diseases. In particular, cell-based immunotherapy has made major strides in the past decade in the treatment of cancer, viral infections, and autoimmune diseases [15–23] (Table 17.1). In August 2017, T cells that have been genetically modified to express tumor-targeting chimeric antigen receptors (CARs) became the first gene therapy to gain approval from the U.S. Food and Drug Administration (FDA) for cancer treatment, highlighting the potential of cellular therapy as a novel treatment option for advanced malignancies.

17.2 Rationale for Cellular Therapies

Cellular therapies – that is, the use of living cells as the therapeutic agent – have a number of distinctive properties that are well suited to the treatment of complex, dynamic diseases. First, mobile living cells are significantly more versatile than single molecules in the type and number of effector functions that can be executed. Cellular therapeutics can be engineered to serve both as independent actors that directly eradicate diseased cells or infectious agents and as payload carriers that deliver therapeutic molecules to a targeted site. For example, cytotoxic T cells expressing surface-bound receptors that direct T cells toward tumor antigens have shown clinical efficacy in treating melanoma [24, 25] and B-cell leukemia [26–29] through direct killing of cancer cells. Antitumor functions can be further enhanced by cellular engineering, such as decorating T-cell surfaces with nanoparticles to specifically deliver drug molecules to the immunological

synapse [30, 31] or stably integrating T cells with DNA constructs that encode for immunostimulatory cytokines under the control of constitutive or inducible promoters [32, 33]. Similarly, genetically engineered stem cells have been programmed to deliver cytotoxic molecules, angiostatic factors, and immunostimulatory cytokines to tumor cells [34–36], demonstrating the versatility and programmability of living cells as therapeutic agents.

Second, unlike static drug molecules, cellular therapeutics can be genetically programmed to conditionally and dynamically deliver functional outputs in response to the presence of specific inputs, thereby increasing therapeutic specificity and efficacy. For example, T cells are naturally programmed to execute functions ranging from cytotoxicity to immune recruitment only upon encountering target cells that express antigens recognized by the T-cell receptor (TCR). T-cell functions vary dynamically with time and are closely coordinated with the rest of the adaptive immune system, thus enabling a finely modulated response to disease and infection. In addition to natural TCRs, synthetic CARs that mimic TCR function and redirect T-cell specificity toward disease targets that are otherwise non-immunogenic have shown great promise in clinical trials [26–29]. Furthermore, T-cell activation can in turn serve as the trigger for downstream effector outputs. For example, by transgenically expressing the immunostimulatory cytokine interleukin-12 (IL-12) gene under the NFAT (nuclear factor of activated T cells) promoter, researchers have generated melanoma-reactive T cells that produce IL-12 only upon T-cell activation, thus avoiding the need for systemic IL-12 injections and associated toxicities [33]. As living entities, therapeutic cells have the ability to perform sense-and-respond functions that greatly enhance treatment specificity and reduce toxic side effects.

Third, unlike chemical pharmaceuticals and biologics, cellular therapeutics have the potential to establish prolonged proliferation in the patient and provide continual surveillance against disease relapse without repeated drug administration. Long-term persistence of therapeutic cells has been shown to be critical in maintaining complete remission across cancer types in adoptive T-cell therapy [37, 38], highlighting the importance of this unique characteristic of cellular therapies.

Despite these important advantages, cellular therapeutics still face major challenges in achieving the level of safety and efficacy required of frontline treatment options. The use of living cells as therapeutic agents invokes a level of complexity not previously seen with traditional pharmaceutical development, and the ability to precisely engineer and stringently regulate therapeutic cells is a critical need that must be fulfilled in the rise of cellular therapy. The following sections discuss some of the challenges facing cell-based therapeutics – particularly cell-based immunotherapies – and highlight solutions that have been developed through the application of synthetic biology.

17.3 Synthetic Biology Approaches to Cellular Immunotherapy Engineering

The programmability of living cells to perform diverse functions – natural or engineered, constitutive or modulated by regulatory systems – is a defining

characteristic and significant advantage of cellular therapies. Viewing cells as chasses, synthetic biologists have demonstrated that biological functions can be rationally designed, systematically optimized, and translated across organisms [39]. A core competency of synthetic biology is the rapid construction, integration, and characterization of biological systems, leading to well-defined, rationally engineered cell products. This approach to cell engineering has generated early examples with potential therapeutic functions and converges with work that has been well established in the field of cellular therapeutics [40–43].

Immune system engineering has played a dominant role in cellular therapies. The application of cell-based immunotherapy can be broadly divided into two categories: immunosuppressive and immunostimulatory. Immunosuppressive therapies aim to dampen aberrant immune responses that characterize inflammatory and autoimmune diseases such as multiple sclerosis, inflammatory bowel diseases, and organ transplant rejection [44, 45]. For example, regulatory T cells and myeloid-derived suppressor cells are naturally immunosuppressive cell types under intensive investigation as treatment options for conditions ranging from ocular inflammation to stroke-induced cerebral ischemia [46, 47]. In contrast, immunostimulatory therapies aim to boost immune responses against infectious agents and tumor growths. Prominent examples in this category include the use of natural killer (NK) cells and cytotoxic T cells that directly kill diseased cells, as well as dendritic cells that stimulate immune responses by presenting disease-associated antigen peptides to effector cells including T cells and B cells [48–50]. The engineering of immune cells provides ample opportunity for synthetic biology to make a real impact on the improvement of health and medicine.

17.3.1 CAR Engineering for Adoptive T-Cell Therapy

Adoptive T-cell therapy is an emerging treatment paradigm in which T cells expressing either TCRs or CARs that target specific disease markers are expanded *ex vivo* prior to infusion into a patient (Figure 17.1). These systemically administered T cells have the ability to seek and destroy target cells that display the cognate antigen, thereby serving as a living drug against otherwise intractable diseases such as refractory cancers and posttransplantation viral infections [51, 52]. In particular, the adoptive transfer of T cells that express anti-CD19 CARs has shown remarkable curative potential against advanced B-cell malignancies, achieving up to 90% complete remission rate in the treatment of acute lymphoblastic leukemia [27, 28, 53].

The development of CARs offers an example of a synthetic biological approach to efficient cell therapy engineering. CARs are synthetic receptors that redirect T-cell specificity toward diseased targets, such as virally infected or cancerous cells, that do not naturally provoke robust immune responses from endogenous T cells. CARs are fusion proteins in which antibody-derived single-chain variable fragments (scFvs) serve as extracellular sensing domains and are fused (via extracellular spacer sequences and transmembrane domains) to cytoplasmic CD3ζ signaling domains derived from the natural TCR [54] (Figure 17.2a). When the CAR is expressed by conventional T cells, ligation of the scFv domain to cognate antigens triggers signaling through the CD3ζ chain, leading to T-cell activation and unleashing cytotoxic

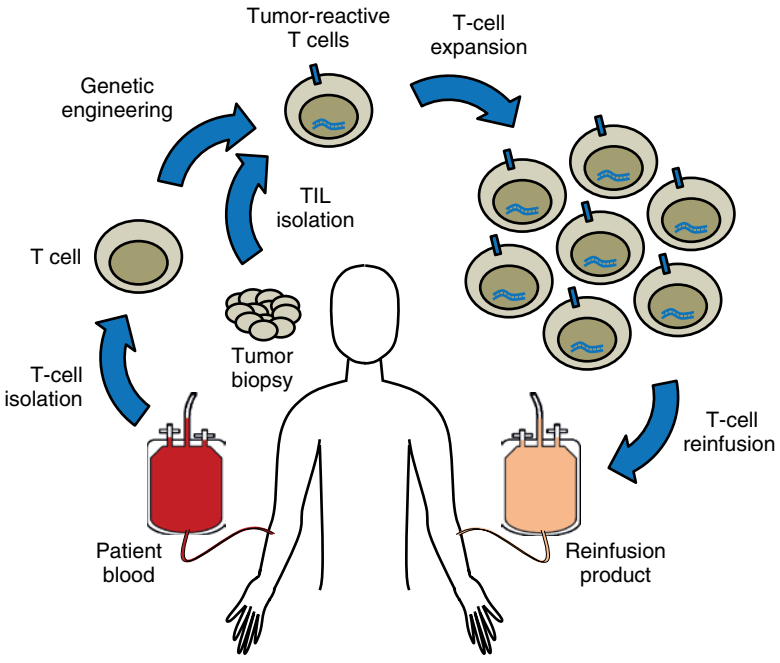
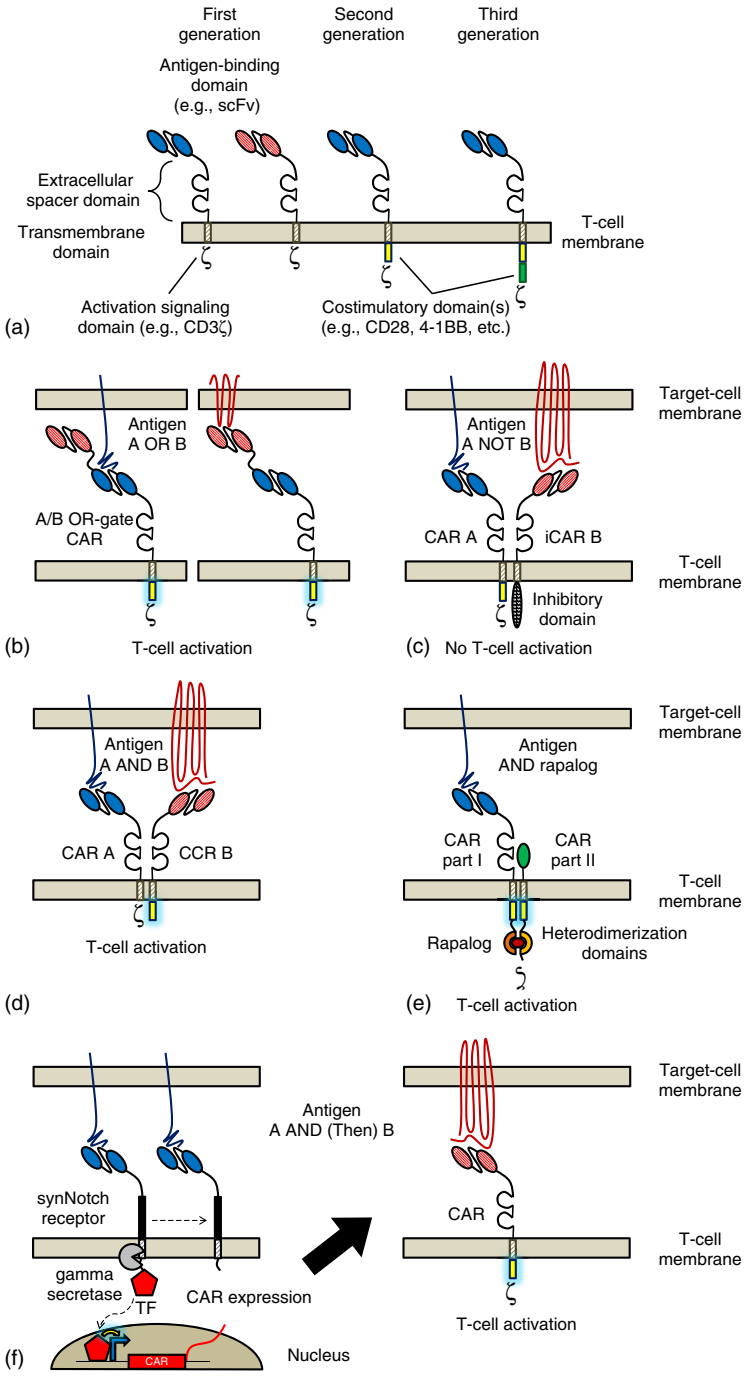


Figure 17.1 Schematic of adoptive T-cell therapy. Endogenous tumor-infiltrating lymphocytes (TILs) are T cells with natural tumor reactivity and can be isolated from tumor biopsies, expanded *ex vivo*, and reinfused into the cancer patient. Alternatively, non-tumor-reactive T cells can be isolated, genetically modified to express a tumor-reactive T-cell receptor (TCR) or chimeric antigen receptor (CAR), expanded *ex vivo*, and reinfused into the patient.

activity toward the target cell. Second- and third-generation CARs have further incorporated costimulatory domains such as CD28 and 4-1BB to enhance T-cell effector functions [55–58]. The modular nature of CARs is highlighted by the diversity of targets that can be recognized by simply replacing the scFv while retaining essentially the same transmembrane and cytoplasmic domains [59].

Although the first CARs predate the emergence of synthetic biology as a discipline, CAR engineering is highly compatible with the synthetic biology approach to biological system design and construction. Taking advantage of the modular composition of CAR molecules, researchers have systematically probed the relationship between CAR structure and T-cell function by characterizing panels of related CAR molecules [60], a process that has been greatly facilitated by the advent of high-throughput DNA synthesis and assembly techniques. For example, studies using a combinatorially constructed panel of CAR molecules have demonstrated that the optimal length of the extracellular spacer in CARs is contingent upon the size of the antigen presented by the target cell [61]. The effects of adding different costimulatory signals and extracellular spacers to CARs have also been explored by combinatorial cloning of CAR molecules [57, 62].

Beyond elucidating principle design rules, this bottom-up approach to CAR engineering has been further leveraged to yield receptors that can execute Boolean logic and regulate T-cell activation according to simultaneous or


First generation
Second generation
Third generation
Antigen-binding domain (e.g., scFv)
Extracellular spacer domain
Transmembrane domain
T-cell membrane
Activation signaling domain (e.g., CD3ζ)
Costimulatory domain(s) (e.g., CD28, 4-1BB, etc.)
(a)
Antigen A OR B
A/B OR-gate CAR
T-cell activation
(b)
Target-cell membrane
Antigen A NOT B
CAR A
iCAR B
T-cell membrane
Inhibitory domain
(c) No T-cell activation
Antigen A AND B
CAR A
CCR B
T-cell activation
(d)
Target-cell membrane
Antigen AND rapalog
CAR part I
CAR part II
T-cell membrane
Rapalog
Heterodimerization domains
(e) T-cell activation
synNotch receptor
gamma secretase
TF
CAR expression
CAR
Nucleus
Antigen A AND (Then) B
Target-cell membrane
CAR
T-cell membrane
T-cell activation
(f)

sequential antigen encounter. Such computational capability offers a means to increase the safety and efficacy of adoptive T-cell therapy by addressing critical clinical challenges, including imperfect targeting specificity and vulnerability to antigen escape (i.e., a process by which diseased cells escape T-cell detection by downregulating the expression of targeted antigens).

For example, three recent clinical trials reported that 38–100% of patient relapses after CD19 CAR-T cell therapy were characterized by the loss of CD19 expression [28, 63, 64]. To address the problem of antigen escape, bispecific OR-gate CARs that incorporate two scFv domains have been developed. T cells armed with OR-gate CARs can respond to either of two distinct antigen inputs, thus reducing the probability that a tumor cell can successfully escape detection via mutational loss of antigen expression [65, 66] (Figure 17.2b). This principle has been applied to generate an optimized CD19/CD20 bispecific CAR that enables cytotoxic T cells to effectively eliminate cancerous B cells that have lost CD19 expression [66, 67]. Specifically, T cells expressing the bispecific CAR are able to not only eradicate established lymphoma in mice but also prevent tumor relapse, whereas animals treated with conventional, single-input CD19 CAR T cells succumb to cancer recurrence caused by antigen escape [66, 67]. Additional combinations such as CD19/CD22 are also under active preclinical evaluation [68], and they promise to significantly increase the efficacy of CAR-T cell therapy against heterogeneous and/or genetically unstable tumors.

Figure 17.2 CARs redirect T-cell specificity toward tumor targets. (a) Schematic of first-, second-, and third-generation CARs. The single-chain variable fragment (scFv) derived from a tumor-antigen-specific antibody serves as the extracellular sensing domain, and the cytoplasmic tail of the CD3ζ chain serves as the intracellular signaling domain of the CAR. In second- and third-generation CARs, one or two costimulatory domains such as CD28 and 4-1BB are directly fused to the CD3ζ chain to enhance T-cell signaling. (b) Schematic of single-chain, bispecific OR-gate CARs. T cells expressing an OR-gate signal processing system can kill any target cell that expresses either antigen A or antigen B. (c) Schematic of an AND-NOT-gate CAR pair. The first receptor is a conventional CAR that targets antigen A. The second is a chimeric inhibitory receptor (iCAR) that targets antigen B and contains the cytoplasmic domain of an inhibitory receptor (e.g., PD-1 or CTLA-4). Presence of antigen A triggers CAR signaling, while presence of antigen B triggers iCAR signaling. The inhibitory function of the iCAR overrides any activation signal that may result from the conventional CAR, thus executing A-NOT-B signal computation. (d) Schematic of an AND-gate CAR pair. The first receptor is a conventional first-generation CAR that targets antigen A and contains only the CD3ζ chain without costimulatory signals. The second is a chimeric costimulatory receptor that targets antigen B and contains both CD28 and 4-1BB costimulatory signals but no CD3ζ chain. Both antigens must be present to trigger a sufficiently robust T-cell response to execute therapeutic function. (e) Schematic of a "remote-controlled" CAR system. Here, the CAR protein is split into two parts, with the first fragment being a conventional CAR that contains the FK506 binding protein (FKBP) instead of the CD3ζ chain at the C-terminus. The second fragment consists of a membrane-tethered CD3ζ chain fused to the FKBP-rapamycin binding (FRB). Presence of a rapamycin analog (rapalog) molecule triggers dimerization between FKBP and FRB, thereby reconstituting a full CAR protein and enabling CAR signaling in response to antigen binding. (f) Schematic of a synthetic Notch (synNotch) receptor-regulated CAR expression system. Upon binding to antigen A, the synNotch receptor releases a TF, which translocates to the nucleus and triggers CAR expression from a cognate promoter. This CAR molecule is subsequently able to trigger T-cell activation upon binding to antigen B, resulting in AND-gate signal computation in a sequential manner.

The modular nature of CAR signaling has also spurred the development of a number of dual-CAR systems that trigger T-cell activation only if two conditions are simultaneously satisfied, in essence executing AND-gate or AND-NOT-gate computations that aim to improve targeting specificity by therapeutic T cells. In one example, researchers designed inhibitory chimeric antigen receptors (iCARs) by replacing the CD3ζ domain of a prostate-specific membrane antigen (PSMA)-targeting CAR with intracellular signaling domains from inhibitory receptors such as cytotoxic T-lymphocyte-associated protein 4 (CTLA-4) and programmed death 1 (PD-1) (Figure 17.2c) [69]. Upon recognition of PSMA, inhibitory signaling through the iCAR effectively competed against activating signaling by a second-generation CD19 CAR to limit T-cell proliferation, cytokine secretion, and cytotoxicity, thereby achieving "CD19-AND-NOT-PSMA" signal computation [69]. It is important to note that NK cells naturally express both activating and inhibitory receptors, and insights to be gained from greater understanding of NK cell signaling may also serve to instruct the robust development of iCARs.

Building on the clinical observation that both the CD3ζ chain and costimulatory signals are necessary to achieve *in vivo* antitumor responses, another study described a dual-receptor system in which the first receptor targets the prostate stem cell antigen (PSCA) and contains only the CD3ζ chain without costimulatory signals, while the second is a chimeric costimulatory receptor that targets PSMA and contains both CD28 and 4-1BB costimulatory signals but no CD3ζ chain (Figure 17.2d). After testing several anti-PSCA scFv domains with varying binding affinities, researchers were able to generate a pair of receptors that trigger T-cell activation and effectively control tumor growth *in vivo* if and only if the tumor expressed both PSCA and PSMA [70]. An interesting alternative approach is to segregate the extracellular scFv from the CD3ζ chain until reconstitution via small molecule-induced heterodimerization. A recent study reported the construction of ON-switch CARs by incorporating the rapamycin analog (rapalog)-inducible heterodimerization domain FK506 binding protein (FKBP) into a truncated second-generation CD19 CAR lacking the CD3ζ chain. Separately, the FKBP-rapamycin binding (FRB) domain was fused to the cytoplasmic portion of CD3ζ (Figure 17.2e) [71]. As such, a fully functional CAR containing the ligand-binding scFv domain and the T-cell-activating CD3ζ chain is only generated upon the addition of rapalog, which induces dimerization between FKBP and FRB, thus bringing the two system components into close proximity. T cells expressing the ON-switch CAR were able to proliferate and mediate cytotoxicity upon target-cell encounter, but only in a rapalog dose-dependent manner, thus yielding temporal control over CAR activation via small molecule drug administration [71].

Yet another example of synthetic receptor design repurposes the signaling mechanism of the Notch receptor. Antigen binding by the Notch receptor exposes a juxtamembrane cleavage sequence that undergoes proteolysis by the intramembrane protease gamma-secretase, a processing step that releases the intracellular Notch domain to the nucleus to serve as a transcription factor (TF) that drives gene expression programs. Utilizing a modular design approach analogous to CAR engineering, researchers developed synthetic Notch (synNotch)

receptors comprised of an extracellular scFv domain fused to a synthetic TF via the endogenous Notch transmembrane domain and juxtamembrane cleavage sequence [72] (Figure 17.2f). Upon ligand binding, the synNotch receptor undergoes cleavage and releases the synthetic TF to drive gene expression from a cognate inducible promoter. By placing CAR expression under this transcriptional control, a dual-receptor system enables T cells to perform AND-gate computation in a sequential manner – that is, antigen A triggers the synNotch receptor to drive expression of the CAR, and subsequent recognition of antigen B by the CAR activates T-cell effector functions. Pairing a green fluorescent protein (GFP) synNotch with a CD19 CAR enables T cells to effectively eliminate tumor cells expressing both GFP and CD19, but not CD19 alone [73].

Although these strategies underscore the vast potential of applied synthetic biology toward enhancing therapeutic efficacy and specificity, CAR performance is still subject to design rules that require better understanding. Multiple recent studies have implicated important CAR components, such as the framework region of scFv domains and the non-signaling extracellular spacer, in triggering tonic signaling [74, 75]. Moreover, in each of the examples highlighted previously, multiple iterations of receptor design were required to identify the correct combination of "modular" components to achieve robust system performance. As more data become available through systematic studies of CAR design parameters, a more quantitative, rational approach to next-generation CAR design will begin to supplant what has largely been a trial-and-error method in engineering CAR-T cells for disease treatment.

17.3.2 Genetic Engineering to Enhance T-Cell Therapeutic Function

Robust proliferation and persistence of T cells have been shown by multiple clinical trials to be both critical to therapeutic efficacy and difficult to achieve *in vivo* [32, 57, 58]. Consequently, there have been many attempts to prolong the survival of CAR-expressing T cells via genetic engineering. These approaches can be broadly grouped into strategies that promote immune stimulation and those that counteract immune suppression. Within the former category, researchers have engineered "armored" T cells to overexpress immunostimulatory cytokines including IL-2, IL-12, and IL-15, thus sustaining T-cell proliferation and effector function [76, 77] (Figure 17.3a). Transgenic expression of costimulatory molecules such as 4-1BB ligand (4-1BBL) and CD40L or surface receptors including interleukin-7 receptor α (IL-7Rα), CCR4, and CXCR2 has also been shown to mitigate T-cell exhaustion and promote T-cell persistence [78–82].

Even when armored with supportive cytokines and costimulatory signaling, engineered T cells can still become exhausted or rendered dysfunctional by repeated antigen stimulation or sustained exposure to immunosuppressive factors located in the tumor microenvironment. To overcome this challenge, extensive research has focused on disrupting endogenous inhibitory signaling pathways (Figure 17.3b) or rewiring immunosuppressive inputs to immunostimulatory outputs (Figure 17.3c). Notably, clinical administration of monoclonal antibodies targeting inhibitory checkpoint molecules such as CTLA-4 and PD-1 has been shown to alleviate immunosuppression of naturally tumor-infiltrating

lymphocytes (TILs) and restore T-cell function in cancer patients [83–85]. While the successes of these treatments have led to US FDA approval of antibodies such as ipilimumab and pembrolizumab, checkpoint blockade therapies can only be effective when tumor-reactive T-cell clones already exist in the patient's system. To address cancer types that are not naturally immunogenic, researchers have engineered tumor-targeting T cells to express dominant-negative receptors (DNRs) that compete with endogenous receptors for binding to tumor-associated

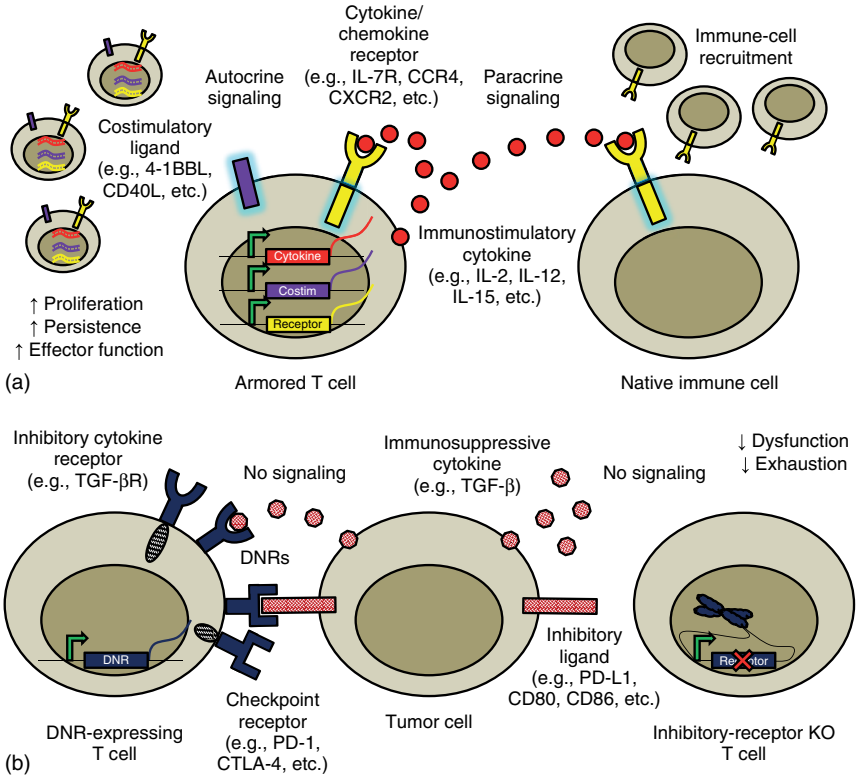


Figure 17.3 Synthetic biological constructs and circuits enable controlled enhancement of T-cell function. (a) "Armored" T cells are engineered to overexpress costimulatory ligands, cytokine receptors, chemokine receptors, or immunostimulatory cytokines that can boost T-cell proliferation, persistence, and effector functions in an autocrine manner. Once secreted, immunostimulatory cytokines (e.g., IL-2, IL-12, IL-15, etc.) can also signal in paracrine fashion to trigger the recruitment, growth, and antitumor responses of native immune cells. (b) T cells can be genetically modified to resist inhibitory signals present on tumor cells (e.g., PD-L1) or within the tumor microenvironment (e.g., TGF-β) by expressing dominant-negative receptors (DNRs) or knocking out inhibitory receptors. DNRs lack signal transduction domains and competitively sequester immunosuppressive ligands away from native inhibitory receptors. Genetic knockout of inhibitory receptor expression abrogates receptor-mediated recognition of immunosuppressive factors, thus reducing T-cell dysfunction and exhaustion. (c) Inverted cytokine receptors (ICRs) are fusions between the extracellular ligand binding of an inhibitory receptor and the intracellular signaling domain of an immunostimulatory receptor. Encounter with immunosuppressive cytokines (e.g., IL-4) in the tumor microenvironment activates expression programs that enhance T-cell proliferation, persistence, and effector functions.

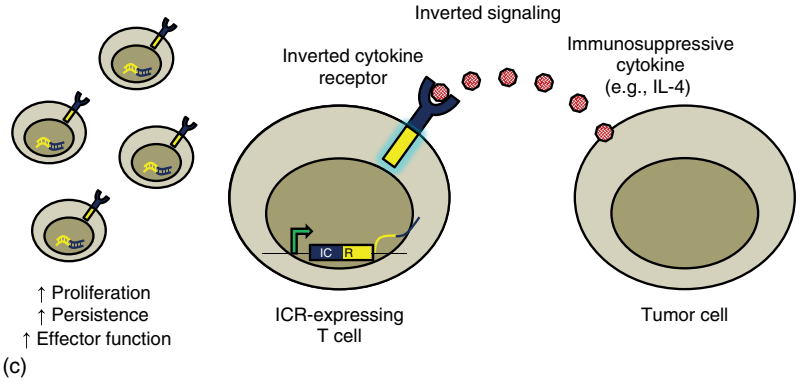


Figure 17.3 (*Continued*)

molecules such as tumor growth factor-beta (TGF-β) or PD-1 (Figure 17.3b) [86–90]. These genetically modified T cells can resist immunosuppression and retain greater *in vivo* antitumor activity and are currently under clinical evaluation (NCT00889954). The rapid progression of clustered regularly interspaced short palindromic repeat (CRISPR)/Cas9 and mammalian genome editing technologies has also enabled the corollary approach of ablating immunosuppressive signaling by knocking out inhibitory receptor expression (Figure 17.3b). Indeed, the first FDA-approved clinical trial involving CRISPR/Cas9-edited cells will examine the use of T cells that had been genetically modified to knockout PD-1 expression [91, 92], and a similar trial has already begun accruing patients abroad (NCT02793856).

To further combat the effect of tumor-mediated immunosuppression, researchers have sought to actively invert suppressive cues to promote T-cell activation. One example of signal inversion was accomplished by fusing the extracellular ligand-binding domain of the inhibitory IL-4 receptor to the intracellular signaling domain of the immunostimulatory IL-7 receptor [93, 94] (Figure 17.3c). Expression of the resulting IL-4/IL-7 inverted cytokine receptor (ICR) reversed the suppression of PSCA-CAR T-cell activity in culture conditions mimicking the tumor milieu of pancreatic cancers [94]. Similarly, another study demonstrated that a chimeric receptor that converts PD-L1 binding to CD28 costimulation could elevate the effector functions of low-avidity T cells to levels observed in high-avidity T cells, suggesting a method to boost the therapeutic efficacy of T cells previously deemed unsuitable for adoptive T-cell therapy [95].

17.3.3 Generating Safer T-Cell Therapeutics with Synthetic Biology

Although extensive research has focused on improving the efficacy of cellular immunotherapy, safety remains the paramount priority in therapeutics development. Even precisely engineered cells retain the possibility of mutation after prolonged periods of expansion inside the host organism. Similarly, sustained

interference with the intricate balance between immunostimulatory and immunosuppressive signaling creates inherent risks for autoimmune dysfunction [96–98]. Safety concerns thus demand gene expression control systems that can be regulated by physician-administered drugs or by molecules specific to the tumor microenvironment. To address this challenge, several ligand-responsive regulatory systems have been developed to control the production of potent cytokines or suicide proteins by engineered T cells [99, 100]. In an early example of mammalian synthetic biology, small molecule-responsive ribozyme switches were inserted in the 3′ untranslated region of transgenes encoding IL-2 and IL-15, resulting in posttranscriptional control of cytokine production in a rapid, reversible manner *in vitro* as well as ligand-dependent modulation of T-cell proliferation *in vivo* [99]. Notably, the ribozyme switch is modularly composed with well-defined sensing, actuating, and information-transmission domains that can be independently modified for the specific application of interest. For example, RNA aptamers to a wide variety of ligands including nucleic acids, small molecules, and proteins have been generated *in vitro* [101], and ribozyme switches tailored for ligands specific to the disease of interest can be rationally designed by incorporating the appropriate RNA aptamers. In the context of cytokine regulation for T-cell therapy, ribozyme switches can be designed to respond to physician-administered drugs or to molecules known to be overexpressed by tumor cells, thus increasing the specificity and safety of this cell-based therapeutic strategy.

As an alternative to the regulation of growth-promoting cytokines, the expression of suicide genes that can rapidly and precisely eliminate engineered T cells provides a means to prevent runaway immune responses. The most commonly used suicide gene is the herpes simplex virus I-derived thymidine kinase (HSV-TK). Originally developed as a method to deplete donor T cells that cause graft-versus-host disease after allogeneic bone marrow transplantation, HSV-TK expression confers sensitivity toward the small molecule drug ganciclovir, thus enabling selective depletion of T cells that have been engineered to transgenically express HSV-TK [102]. However, HSV-TK-mediated cell depletion is often incomplete, and the strategy precludes the use of ganciclovir as an antiviral drug for cytomegalovirus infections, a common and often fatal complication of bone marrow transplants [103]. Taking an alternative approach, researchers have developed chimeric suicide genes that fuse pro-apoptotic proteins with dimerization domains to induce apoptosis upon the administration of a chemical ligand [104]. For example, an inducible caspase 9 suicide system has been constructed by fusing an inactive pro-caspase 9 monomer to FKBP [105]. Upon administration of the chemical inducer of dimerization (CID) molecule AP1903, the FKBP domains homodimerize, resulting in the cross-linking and activation of the tethered caspase 9 domains, which in turn induce apoptosis in cells expressing this suicide system (Figure 17.4). The inducible caspase 9 system has been tested in an adoptive T-cell therapy trial and demonstrated the ability to eradicate >90% of engineered T cells within 30 min of AP1903 administration, effectively eliminating graft-versus-host disease symptoms without recurrence [100]. Combining several of the technologies summarized previously, researchers have engineered second-generation CD19 CAR-T cells equipped with

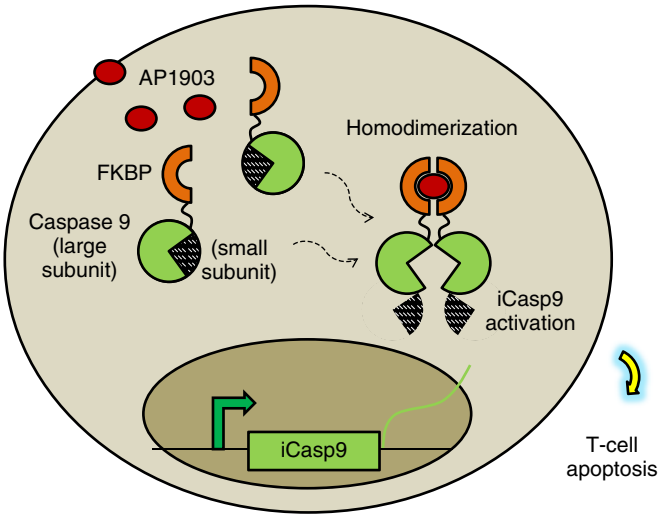


Figure 17.4 A chemically inducible caspase 9 kill switch. Inactive pro-caspase 9 monomers are linked to the human FK506 binding protein FKBP and constitutively expressed in the engineered cell. Upon addition of the chemical inducer of dimerization AP1903, the FKBP domains dimerize and lead to the cross-linking and activation of caspase 9, which triggers downstream events in the apoptosis pathway and results in cell death.

constitutive IL-15 production and the inducible caspase 9 suicide system and demonstrated superior *in vivo* T-cell expansion and antitumor effects compared with T cells expressing the CAR alone [106].

Although suicide gene systems provide a powerful countermeasure to major adverse events such as deleterious genetic mutations in engineered cells, results from clinical trials have also highlighted situations in which measured dampening of functions rather than complete elimination of therapeutic cells is the preferred response. In adoptive T-cell therapy for cancer, tumor regression is strongly associated with a dramatic increase in the level of inflammatory cytokines, a phenomenon known as cytokine storm or tumor lysis syndrome [107, 108]. When the intensity of the tumor lysis syndrome exceeds physiological tolerance, corticosteroids can be administered to the patient, thereby not only quelling the immediate dangers of therapy-associated toxicity but also terminating the treatment by effectively disabling the therapeutic cell population [109]. As a potential alternative, researchers have engineered synthetic circuits that regulate the amplitude of T-cell activation, thus enabling fine-tuning of T-cell-mediated responses [110]. The bacterial protein OspF downregulates T-cell activation by inactivating the extracellular signal-regulated kinase (ERK). An "amplitude limiter" consisting of a negative feedback loop with OspF expressed from an NFAT promoter lowers the maximum level of T-cell activation-induced gene expression, which can be further modulated by the addition of degradation tags to the OspF protein. Furthermore, a "pause switch" was constructed by expressing OspF from a doxycycline-inducible promoter, such that pulses of doxycycline addition result in temporary reductions in T-cell activation-induced expression [110].

In addition to the magnitude of the immune response, the precision of effector functions is critical to the development of safer cell-based immunotherapy. For example, there is an emerging consensus that the lack of tumor-exclusive, surface-bound antigens presents a fundamental challenge to the widespread implementation of adoptive T-cell therapy [111, 112]. T cells identify target cells via surface receptor-mediated recognition of membrane-bound biomarker. However, tumor cells rarely express surface antigens that are completely absent in all healthy tissues. As a result, basal antigen expression by healthy tissues frequently elicits "on-target, off-tumor" toxicities in adoptive T-cell therapy. Bispecific CAR-T cells capable of AND- or AND-NOT-gate signal computation discussed previously represent one approach to increasing the precision of disease-cell recognition based on surface interactions [69–73]. Another recently reported approach endows T cells with the ability to interrogate the *intracellular* environment of target cells through the delivery of intracellular antigen-responsive cytotoxic switches derived from the cytotoxic proteins granzyme B (GrB) [113]. Expression of a small ubiquitin-like modifier (SUMO)–GrB fusion protein selectively triggered cytotoxicity in cells overexpressing the intracellular tumor-associated sentrin-specific protease 1 (SENP1) [113]. Coupled with demonstrations of recombinant GrB transfer from T cells into target cells, these results point to a potentially viable strategy for improving the therapeutic precision of adoptive T-cell therapy by expanding the repertoire of targetable candidate antigens to include a plethora of intracellular disease signatures. Although these systems remain to be validated *in vivo*, they highlight the versatility and malleability of cellular therapeutics, as well as the importance of effective engineering techniques in the development and optimization of cell-based immunotherapy.

17.4 Challenges and Future Outlook

The ability to efficiently design, construct, and optimize synthetic biological systems that modify and/or interface with living cells is expanding new possibilities in the development of cellular therapeutics and offering enticing views of next-generation strategies for disease treatment. Early synthetic biological circuits consist of input/output devices linked in various configurations to achieve diverse purposes, including signal oscillation, memory, and cell–cell communication [114–117]. If engineered to fit the clinical context and application-specific requirements, such functions could significantly improve the performance of cellular therapeutics. For example, a robust, tunable oscillation pattern would enable the regular, pulsatory delivery of drug molecules that are either synthesized or carried by therapeutic cells. The ability to memorize and keep count of events such as cell divisions would enable timed proliferation and death of engineered cells, providing an additional mechanism to ensure the safety of cellular therapies. The ability to sense extracellular molecular signals and communicate with other cells could enable time-, position-, and community-dependent responses that serve as disease diagnostics or enhance the specificity of cellular therapeutics toward disease targets. In addition to synthetic circuitry that confers novel functions onto engineered cells, rapidly advancing genome editing

technologies are poised to enable the generation of universal, off-the-shelf cellular therapeutics that lack antigenic markers to induce immune rejection, a development that would significantly reduce the time and financial costs associated with producing personalized supplies of therapeutic cells for each patient [118]. If efforts in whole-genome synthesis and the construction of artificial cells come to fruition, cellular therapeutics may eventually consist of fully synthetic cells with precisely controlled functions.

Despite the myriad possibilities that synthetic biology inspires, real obstacles need to be overcome in moving from model systems to real-world applications in health and medicine. First, most synthetic biological systems demonstrated to date have been designed to function in microorganisms such as yeasts and bacteria rather than mammalian cells. Although some studies have shown transportability across organisms [99, 119], significantly more experience will be required in mammalian cell engineering to achieve the level of efficiency in system assembly, integration, and characterization that is now possible in microorganisms.

Second, despite the variety of synthetic circuits that have been reported, a relatively small number of parts (e.g., the tet-inducible promoter, the theophylline aptamer, fluorescent protein outputs, or acyl-homoserine lactone (AHL)-based quorum sensing components) have been reused in a large number of designs, reflecting a need to expand the inventory of biological parts. In particular, cellular therapeutics development will require new outputs that execute therapeutic functions at precisely defined activity levels [120], a significantly more complex task than ON/OFF control of fluorescent protein outputs. Similarly, new sensors need to be developed to respond to therapeutically relevant inputs such as disease-associated metabolites or FDA-approved drugs rather than oft-used but clinically unacceptable inputs such as theophylline or isopropyl β-D-1-thiogalactopyranoside (IPTG).

Third, given the paramount importance of safety in medical applications, any synthetic system applied to cellular therapeutics must perform with consistency and precision in the face of heterogeneities that are inevitable in the human body and particularly in diseased cells. Unlike model systems in which parameters such as input ligand concentration and cell density can be precisely controlled, clinical applications in which heterogeneous cell populations harvested from patients need to be quickly genetically modified, expanded, and reinfused in bulk into the patients require a high level of robustness such that the system would generate predictable and consistent outputs without the benefit of extensive cell-population refinement or well-defined ranges of input signal strength. In this regard, researchers are actively investigating genetic engineering strategies that can enable synthetic components to interface more robustly with host cell physiology. In contrast to random insertion of CAR transgenes via viral transduction, site-specific integration of a CD19 CAR into the T-cell receptor α chain (TRAC) locus of primary human T cells resulted in antigen-stimulated regulation and greater uniformity of CAR expression [121, 122]. These site-specifically modified T cells exhibited reduced tonic signaling, delayed T-cell exhaustion, and enhanced antitumor potency [121], underscoring the importance of being able to tune the expression level and signaling strength of synthetic systems.

Finally, as synthetic biologists build increasingly complex systems, the twin issues of scaling and implementation must be addressed. Current strategies in circuit design generally result in a roughly linear relationship between part number/size and functionality. For example, an RNAi-based cancer-cell identifier has been demonstrated to distinguish HeLa cells from a number of other cancer cell lines by sensing the levels of six distinct microRNAs (miRNAs) through a network of constitutive and inducible promoters linked to genes encoding various inducer proteins and miRNA target sites [123]. This work provided an elegant example of a synthetic, multi-input system in mammalian cells applicable to cellular therapeutics engineering. However, adding each new miRNA input would require a significant increase in the footprint and complexity of the computation network without the benefit of economy of scale, leading to problems of parts shortage (i.e., there are limited numbers of inducible promoters available) and low integration efficiency (i.e., the system will eventually be too large to be delivered and stably integrated in the cell). This challenge of scalability will have to be resolved before a broad range of cellular therapeutics with the necessary level of functional complexity can be developed through synthetic biology.

Cellular therapeutics has generated a tremendous amount of excitement in recent years, particularly in cancer immunotherapy. The cell engineering techniques and circuit design expertise being developed through synthetic biology are poised to make timely and significant contributions to the continuing improvement of cellular therapeutics. Important challenges remain to be addressed in biological system design and implementation methods, and the accumulating knowledge from ongoing efforts in synthetic biology will be critical in the construction of synthetic biological systems with real-world applications in health and medicine.

Acknowledgment

This material is based in part upon work supported by the National Institutes of Health (DP5OD012133-01 and P50CA092131) and the National Science Foundation (CBET 1533767). P.H. is supported by the NIH Biotechnology Training in Biomedical Sciences and Engineering Program (T32 GM067555).

Definitions

Cellular therapy The use of living cells, as distinct from chemical pharmaceuticals or biologics, as therapeutic agents in the treatment of diseases

Immunotherapy Disease treatment that modulates the immune system to enhance immune responses against disease agents or diseased cells

Adoptive T-cell therapy A type of cellular immunotherapy in which autologous T cells with specificity toward disease targets, including cancerous and virally infected cells, are expanded *ex vivo* and reinfused into the patient. T cells harvested from the patient may have endogenous disease-specific

reactivity, or they may be genetically modified to express disease-targeting receptors prior to expansion and reinfusion

Chimeric antigen receptor (CAR) A class of membrane-bound fusion proteins that redirect T-cell specificity toward specified target antigens. First-generation CARs consist of four major domains: an extracellular single-chain variable fragment (scFv) that determines target specificity, an extracellular spacer typically derived from immunoglobulin molecules, a transmembrane domain, and the cytoplasmic signaling domain of the CD3ζ chain that triggers T-cell activation upon ligand binding to the scFv. Second- and third-generation CARs contain one or two additional cytoplasmic costimulatory domains, respectively, that enhance T-cell activation. The most commonly used costimulatory domains are CD28 and 4-1BB

Suicide gene Genes encoding for protein products that lead to cell death either directly (e.g., by triggering the apoptosis pathway) or indirectly (e.g., by processing prodrugs to lethal products through enzymatic reactions)

References

1 Lienert, F., Lohmueller, J.J., Garg, A., and Silver, P.A. (2014) Synthetic biology in mammalian cells: next generation research tools and therapeutics. *Nat. Rev. Mol. Cell Biol.*, **15** (2), 95–107.

2 Kojima, R., Aubel, D., and Fussenegger, M. (2016) Toward a world of theranostic medication: programming biological sentinel systems for therapeutic intervention. *Adv. Drug Delivery Rev.*, **105** (Pt A), 66–76.

3 Dudek, R.M., Chuang, Y., and Leonard, J.N. (2014) *Engineered Cell-Based Therapies: A Vanguard of Design-Driven Medicine*, Springer, New York, pp. 369–391.

4 Fischbach, M.A., Bluestone, J.A., and Lim, W.A. (2013) Cell-based therapeutics: the next pillar of medicine. *Sci. Transl. Med.*, **5** (179), 179ps7.

5 Adams, G.P. and Weiner, L.M. (2005) Monoclonal antibody therapy of cancer. *Nat. Biotechnol.*, **23** (9), 1147–1157.

6 Hudis, C.A. (2007) Trastuzumab – mechanism of action and use in clinical practice. *N. Engl. J. Med.*, **357** (1), 39–51.

7 Genentech (1978) First Successful Laboratory Production of Human Insulin Announced.

8 Egrie, J.C., Strickland, T.W., Lane, J. *et al.* (1986) Characterization and biological effects of recombinant human erythropoietin. *Immunobiology*, **172** (3–5), 213–224.

9 *Cancer Facts & Figures* (2016). Atlanta: American Cancer Society.

10 Johnson, D.R. and O'Neill, B.P. (2012) Glioblastoma survival in the United States before and during the temozolomide era. *J. Neuro-Oncol.*, **107** (2), 359–364.

11 Patil, C.G., Yi, A., Elramsisy, A. *et al.* (2012) Prognosis of patients with multifocal glioblastoma: a case–control study. *J. Neurosurg.*, **117** (4), 705–711.

12 Greenman, C., Stephens, P., Smith, R. *et al.* (2007) Patterns of somatic mutation in human cancer genomes. *Nature*, **446** (7132), 153–158.

13 Bovenberg, M.S.S., Degeling, M.H., and Tannous, B.A. (2013) Advances in stem cell therapy against gliomas. *Trends Mol. Med.*, **19** (5), 281–291.

14 Boulikas, T. and Vougiouka, M. (2004) Recent clinical trials using cisplatin, carboplatin and their combination chemotherapy drugs (review). *Oncol. Rep.*, **11** (3), 559–595.

15 Chang, Z.L. and Chen, Y.Y. (2017) CARs: synthetic immunoreceptors for cancer therapy and beyond. *Trends Mol. Med.*, **23** (5), 430–450.

16 Deeks, S.G., Wagner, B., Anton, P.A. *et al.* (2002) A phase II randomized study of HIV-specific T-cell gene therapy in subjects with undetectable plasma viremia on combination antiretroviral therapy. *Mol. Ther.*, **5** (6), 788–797.

17 Fransson, M., Piras, E., Burman, J. *et al.* (2012) CAR/FoxP3-engineered T regulatory cells target the CNS and suppress EAE upon intranasal delivery. *J. Neuroinflammation*, **9** (1), 112.

18 MacDonald, K.G., Hoeppli, R.E., Huang, Q. *et al.* (2016) Alloantigen-specific regulatory T cells generated with a chimeric antigen receptor. *J. Clin. Invest.*, **126** (4), 1413–1424.

19 Czaja, A.J. (2015) Adoptive cell transfer in autoimmune hepatitis. *Expert Rev. Gastroenterol. Hepatol.*, **9** (6), 1–16.

20 Chodon, T., Comin-Anduix, B., Chmielowski, B. *et al.* (2014) Adoptive transfer of MART-1 T-cell receptor transgenic lymphocytes and dendritic cell vaccination in patients with metastatic melanoma. *Clin. Cancer Res.*, **20** (9), 2457–2465.

21 Yang, L., Yang, H., Rideout, K. *et al.* (2008) Engineered lentivector targeting of dendritic cells for in vivo immunization. *Nat. Biotechnol.*, **26** (3), 326–334.

22 Guillerey, C., Huntington, N.D., and Smyth, M.J. (2016) Targeting natural killer cells in cancer immunotherapy. *Nat. Immunol.*, **17** (9), 1025–1036.

23 Rezvani, K. and Rouce, R.H. (2015) The application of natural killer cell immunotherapy for the treatment of cancer. *Front. Immunol.*, **6**, 578.

24 Lu, Y.-C., Yao, X., Crystal, J.S. *et al.* (2014) Efficient identification of mutated cancer antigens recognized by T cells associated with durable tumor regressions. *Clin. Cancer Res.*, **20** (13), 3401–3410.

25 Robbins, P.F., Lu, Y.-C., El-Gamil, M. *et al.* (2013) Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nat. Med.*, **19** (6), 747–752.

26 Kochenderfer, J.N., Dudley, M.E., Kassim, S.H. *et al.* (2015) Chemotherapy-refractory diffuse large B-cell lymphoma and indolent B-cell malignancies can be effectively treated with autologous T cells expressing an anti-CD19 chimeric antigen receptor. *J. Clin. Oncol.*, **33**, 540–549.

27 Davila, M.L., Riviere, I., Wang, X. *et al.* (2014) Efficacy and toxicity management of 19-28z CAR T cell therapy in B cell acute lymphoblastic leukemia. *Sci. Transl. Med.*, **6** (224), 224ra225.

28 Maude, S.L., Frey, N., Shaw, P.A. *et al.* (2014) Chimeric antigen receptor T cells for sustained remissions in leukemia. *N. Engl. J. Med.*, **371** (16), 1507–1517.

29 Wang, X., Popplewell, L.L., Wagner, J.R. *et al.* (2016) Phase 1 studies of central memory–derived CD19 CAR T–cell therapy following autologous HSCT in patients with B-cell NHL. *Blood*, **127** (24), 2980–2990.

30 Stephan, M.T., Stephan, S.B., Bak, P. *et al.* (2012) Synapse-directed delivery of immunomodulators using T-cell-conjugated nanoparticles. *Biomaterials*, **33** (23), 5776–5787.

31 Jones, R.B., Mueller, S., Kumari, S. *et al.* (2017) Antigen recognition-triggered drug delivery mediated by nanocapsule-functionalized cytotoxic T-cells. *Biomaterials*, **117**, 44–53.

32 Pegram, H.J., Lee, J.C., Hayman, E.G. *et al.* (2012) Tumor-targeted T cells modified to secrete IL-12 eradicate systemic tumors without need for prior conditioning. *Blood*, **119** (18), 4133–4141.

33 Zhang, L., Kerkar, S.P., Yu, Z. *et al.* (2011) Improving adoptive T cell therapy by targeting and controlling IL-12 expression to the tumor environment. *Mol. Ther.*, **19** (4), 751–759.

34 Hingtgen, S.D., Kasmieh, R., van de Water, J. *et al.* (2010) A novel molecule integrating therapeutic and diagnostic activities reveals multiple aspects of stem cell-based therapy. *Stem Cells*, **28** (4), 832–841.

35 Yin, J., Kim, J.-K., Moon, J.-H. *et al.* (2011) hMSC-mediated concurrent delivery of endostatin and carboxylesterase to mouse xenografts suppresses glioma initiation and recurrence. *Mol. Ther.*, **19** (6), 1161–1169.

36 Ryu, C.H., Park, S.-H., Park, S.A. *et al.* (2011) Gene therapy of intracranial glioma using interleukin 12-secreting human umbilical cord blood-derived mesenchymal stem cells. *Hum. Gene Ther.*, **22** (6), 733–743.

37 Leen, A.M., Rooney, C.M., and Foster, A.E. (2007) Improving T cell therapy for cancer. *Annu. Rev. Immunol.*, **25** (1), 243–265.

38 Robbins, P.F., Dudley, M.E., Wunderlich, J. *et al.* (2004) Cutting edge: persistence of transferred lymphocyte clonotypes correlates with cancer regression in patients receiving cell transfer therapy. *J. Immunol.*, **173** (12), 7125–7130.

39 Chen, Y.Y., Galloway, K.E., and Smolke, C.D. (2012) Synthetic biology: advancing biological frontiers by building synthetic systems. *Genome Biol.*, **13** (2), 240.

40 Kalos, M. and June, C.H. (2013) Adoptive T cell transfer for cancer immunotherapy in the era of synthetic biology. *Immunity*, **39** (1), 49–60.

41 Chakravarti, D. and Wong, W.W. (2015) Synthetic biology in cell-based cancer immunotherapy. *Trends Biotechnol.*, **33** (8), 449–461.

42 Geering, B. and Fussenegger, M. (2015) Synthetic immunology: modulating the human immune system. *Trends Biotechnol.*, **33** (2), 65–79.

43 Lim, W.A. and June, C.H. (2017) The principles of engineering immune cells to treat cancer. *Cell*, **168** (4), 724–740.

44 Safinia, N., Leech, J., Hernandez-Fuentes, M. *et al.* (2013) Promoting transplantation tolerance; adoptive regulatory T cell therapy. *Clin. Exp. Immunol.*, **172** (2), 158–168.

45 Bluestone, J.A. (2005) Regulatory T-cell therapy: is it ready for the clinic? *Nat. Rev. Immunol.*, **5** (4), 343–349.

46 Forrester, J.V., Steptoe, R.J., Klaska, I.P. *et al.* (2013) Cell-based therapies for ocular inflammation. *Prog. Retin. Eye Res.*, **35**, 82–101.

47 Li, P., Gan, Y., Sun, B.-L. *et al.* (2013) Adoptive regulatory T-cell therapy protects against cerebral ischemia. *Ann. Neurol.*, **74** (3), 458–471.

48 Schürch, C.M., Riether, C., and Ochsenbein, A.F. (2013) Dendritic cell-based immunotherapy for myeloid leukemias. *Front. Immunol.*, **4**, 496.

49 Cheng, M., Chen, Y., Xiao, W. *et al.* (2013) NK cell-based immunotherapy for malignant diseases. *Cell. Mol. Immunol.*, **10** (3), 230–252.

50 Hildebrandt, M., Peggs, K., Uharek, L. *et al.* (2014) Immunotherapy: opportunities, risks and future perspectives. *Cytotherapy*, **16** (4 Suppl), S120–S129.

51 Rosenberg, S.A. (2012) Raising the bar: the curative potential of human cancer immunotherapy. *Sci. Transl. Med.*, **4** (127), 127ps8.

52 Full, F., Lehner, M., Thonn, V. *et al.* (2010) T cells engineered with a cytomegalovirus-specific chimeric immunoreceptor. *J. Virol.*, **84** (8), 4083–4088.

53 Brentjens, R.J., Davila, M.L., Riviere, I. *et al.* (2013) CD19-targeted T cells rapidly induce molecular remissions in adults with chemotherapy-refractory acute lymphoblastic leukemia. *Sci. Transl. Med.*, **5** (177), 177ra38.

54 Eshhar, Z., Waks, T., Gross, G., and Schindler, D.G. (1993) Specific activation and targeting of cytotoxic lymphocytes through chimeric single chains consisting of antibody-binding domains and the gamma or zeta subunits of the immunoglobulin and T-cell receptors. *Proc. Natl. Acad. Sci. U.S.A.*, **90** (2), 720–724.

55 Imai, C., Mihara, K., Andreansky, M. *et al.* (2004) Chimeric receptors with 4-1BB signaling capacity provoke potent cytotoxicity against acute lymphoblastic leukemia. *Leukemia*, **18** (4), 676–684.

56 Kowolik, C.M., Topp, M.S., Gonzalez, S. *et al.* (2006) CD28 costimulation provided through a CD19-specific chimeric antigen receptor enhances in vivo persistence and antitumor efficacy of adoptively transferred T cells. *Cancer Res.*, **66** (22), 10995–11004.

57 Pulè, M.A., Straathof, K.C., Dotti, G. *et al.* (2005) A chimeric T cell antigen receptor that augments cytokine release and supports clonal expansion of primary human T cells. *Mol. Ther.*, **12** (5), 933–941.

58 Song, D.-G., Ye, Q., Poussin, M. *et al.* (2012) CD27 costimulation augments the survival and antitumor activity of redirected human T cells in vivo. *Blood*, **119** (3), 696–706.

59 Sadelain, M., Brentjens, R., and Rivière, I. (2009) The promise and potential pitfalls of chimeric antigen receptors. *Curr. Opin. Immunol.*, **21** (2), 215–223.

60 Maher, J., Brentjens, R.J., Gunset, G. *et al.* (2002) Human T-lymphocyte cytotoxicity and proliferation directed by a single chimeric TCRzeta/CD28 receptor. *Nat. Biotechnol.*, **20** (1), 70–75.

61 Hudecek, M., Lupo-Stanghellini, M.-T., Kosasih, P.L. *et al.* (2013) Receptor affinity and extracellular domain modifications affect tumor recognition by ROR1-specific chimeric antigen receptor T cells. *Clin. Cancer Res.*, **19** (12), 3153–3164.

62 Carpenito, C., Milone, M.C., Hassan, R. *et al.* (2009) Control of large, established tumor xenografts with genetically retargeted human T cells containing CD28 and CD137 domains. *Proc. Natl. Acad. Sci. U.S.A.*, **106** (9), 3360–3365.

63 Lee, D.W., Kochenderfer, J.N., Stetler-Stevenson, M. *et al.* (2015) T cells expressing CD19 chimeric antigen receptors for acute lymphoblastic leukaemia in children and young adults: a phase 1 dose-escalation trial. *Lancet (London, England)*, **385** (9967), 517–528.

64 Gardner, R.A., Finney, O., Annesley, C. *et al.* (2017) Intent to treat leukemia remission by CD19CAR T cells of defined formulation and dose in children and young adults. *Blood*. doi: 10.1182/blood-2017-02-769208

65 Grada, Z., Hegde, M., Byrd, T. *et al.* (2013) TanCAR: a novel bispecific chimeric antigen receptor for cancer immunotherapy. *Mol. Ther. Nucleic Acids*, **2**, e105.

66 Zah, E., Lin, M.-Y., Silva-Benedict, A. *et al.* (2016) T cells expressing CD19/CD20 bi-specific chimeric antigen receptors prevent antigen escape by malignant B cells. *Cancer Immunol. Res.*, **4** (6), 498–508.

67 Zah, E., Lin, M.-Y., Silva-Benedict, A. *et al.* (2016) ADDENDUM: T cells expressing CD19/CD20 bispecific chimeric antigen receptors prevent antigen escape by malignant B cells. *Cancer Immunol. Res.*, **4** (7), 639–641.

68 Qin, H., Haso, W., Nguyen, S.M., and Fry, T.J. (2015) Preclinical development of bispecific chimeric antigen receptor targeting both CD19 and CD22. *Blood*, **126** (23).

69 Fedorov, V.D., Themeli, M., and Sadelain, M. (2013) PD-1- and CTLA-4-based inhibitory chimeric antigen receptors (iCARs) divert off-target immunotherapy responses. *Sci. Transl. Med.*, **5** (215), 215ra172.

70 Kloss, C.C., Condomines, M., Cartellieri, M. *et al.* (2013) Combinatorial antigen recognition with balanced signaling promotes selective tumor eradication by engineered T cells. *Nat. Biotechnol.*, **31** (1), 71–75.

71 Wu, C.-Y., Roybal, K.T., Puchner, E.M. *et al.* (2015) Remote control of therapeutic T cells through a small molecule-gated chimeric receptor. *Science*, **350** (6258), aab4077.

72 Morsut, L., Roybal, K.T., Xiong, X. *et al.* (2016) Engineering customized cell sensing and response behaviors using synthetic notch receptors. *Cell*, **164** (4), 780–791.

73 Roybal, K.T., Rupp, L.J., Morsut, L. *et al.* (2016) Precision tumor recognition by T cells with combinatorial antigen-sensing circuits. *Cell*, **164** (4), 770–779.

74 Long, A.H., Haso, W.M., Shern, J.F. *et al.* (2015) 4-1BB costimulation ameliorates T cell exhaustion induced by tonic signaling of chimeric antigen receptors. *Nat. Med.*, **21** (6), 581–590.

75 Watanabe, N., Bajgain, P., Sukumaran, S. *et al.* (2016) Fine-tuning the CAR spacer improves T-cell potency. *Oncoimmunology*, **5** (12), e1253656.

76 Pegram, H.J., Park, J.H., and Brentjens, R.J. (2014) CD28z CARs and armored CARs. *Cancer J.*, **20** (2), 127–133.

77 Chmielewski, M. and Abken, H. (2015) TRUCKs: the fourth generation of CARs. *Expert Opin. Biol. Ther.*, **15** (8), 1145–1154.

78 Yeku, O.O. and Brentjens, R.J. (2016) Armored CAR T-cells: utilizing cytokines and pro-inflammatory ligands to enhance CAR T-cell anti-tumour efficacy. *Biochem. Soc. Trans.*, **44** (2), 412–418.

79 Perna, S.K., Pagliara, D., Mahendravada, A. *et al.* (2014) Interleukin-7 mediates selective expansion of tumor-redirected cytotoxic T lymphocytes (CTLs) without enhancement of regulatory T-cell inhibition. *Clin. Cancer Res.*, **20** (1), 131–139.

80 Di Stasi, A., De Angelis, B., Rooney, C.M. *et al.* (2009) T lymphocytes coexpressing CCR4 and a chimeric antigen receptor targeting CD30 have

improved homing and antitumor activity in a Hodgkin tumor model. *Blood*, **113** (25), 6392–6402.

81 Peng, W., Ye, Y., Rabinovich, B.A. *et al.* (2010) Transduction of tumor-specific T cells with CXCR2 chemokine receptor improves migration to tumor and antitumor immune responses. *Clin. Cancer Res.*, **16** (22), 5458–5468.

82 Zhao, Z., Condomines, M., van der Stegen, S.J.C. *et al.* (2015) Structural design of engineered costimulation determines tumor rejection kinetics and persistence of CAR T cells. *Cancer Cell*, **28** (4), 415–428.

83 Peggs, K.S. and Quezada, S.A. (2010) Ipilimumab: attenuation of an inhibitory immune checkpoint improves survival in metastatic melanoma. *Expert Rev. Anticancer Ther.*, **10** (11), 1697–1701.

84 Callahan, M.K. and Wolchok, J.D. (2013) At the bedside: CTLA-4- and PD-1-blocking antibodies in cancer immunotherapy. *J. Leukocyte Biol.*, **94** (1), 41–53.

85 Shin, D.S. and Ribas, A. (2015) The evolution of checkpoint blockade as a cancer therapy: what's here, what's next? *Curr. Opin. Immunol.*, **33**, 23–35.

86 Gorelik, L. and Flavell, R.A. (2001) Immune-mediated eradication of tumors through the blockade of transforming growth factor-beta signaling in T cells. *Nat. Med.*, **7** (10), 1118–1122.

87 Bendle, G.M., Linnemann, C., Bies, L. *et al.* (2013) Blockade of TGF-β signaling greatly enhances the efficacy of TCR gene therapy of cancer. *J. Immunol.*, **191** (6), 3232–3239.

88 Foster, A.E., Dotti, G., Lu, A. *et al.* (2008) Antitumor activity of EBV-specific T lymphocytes transduced with a dominant negative TGF-beta receptor. *J. Immunother.*, **31** (5), 500–505.

89 Shin, J.H., Park, H.B., and Choi, K. (2016) Enhanced anti-tumor reactivity of cytotoxic T lymphocytes expressing PD-1 decoy. *Immune Netw.*, **16** (2), 134.

90 Cherkassky, L., Morello, A., Villena-Vargas, J. *et al.* (2016) Human CAR T cells with cell-intrinsic PD-1 checkpoint blockade resist tumor-mediated inhibition. *J. Clin. Invest.*, **126** (8), 3130–3144.

91 Ren, J., Zhang, X., Liu, X. *et al.* (2017) A versatile system for rapid multiplex genome-edited CAR T cell generation. *Oncotarget*, **8** (10), 17002–17011.

92 Rupp, L.J., Schumann, K., Roybal, K.T. *et al.* (2017) CRISPR/Cas9-mediated PD-1 disruption enhances anti-tumor efficacy of human chimeric antigen receptor T cells. *Sci. Rep.*, **7** (1), 737.

93 Leen, A.M., Sukumaran, S., Watanabe, N. *et al.* (2014) Reversal of tumor immune inhibition using a chimeric cytokine receptor. *Mol. Ther.*, **22** (6), 1211–1220.

94 Mohammed, S., Sukumaran, S., Bajgain, P. *et al.* (2017) Improving chimeric antigen receptor-modified T cell function by reversing the immunosuppressive tumor microenvironment of pancreatic cancer. *Mol. Ther.*, **25** (1), 249–258.

95 Schlenker, R., Olguín-Contreras, L.F., Leisegang, M. *et al.* (2017) Chimeric PD-1:28 receptor upgrades low-avidity T cells and restores effector function of tumor-infiltrating lymphocytes for adoptive cell therapy. *Cancer Res.*, **77**, 3577–3590.

96 Zhang, L., Morgan, R.A., Beane, J.D. *et al.* (2015) Tumor-infiltrating lymphocytes genetically engineered with an inducible gene encoding interleukin-12 for the immunotherapy of metastatic melanoma. *Clin. Cancer Res.*, **21** (10), 2278–2288.

97 Michot, J.M., Bigenwald, C., Champiat, S. *et al.* (2016) Immune-related adverse events with immune checkpoint blockade: a comprehensive review. *Eur. J. Cancer*, **54**, 139–148.

98 Naidoo, J., Page, D.B., Li, B.T. *et al.* (2015) Toxicities of the anti-PD-1 and anti-PD-L1 immune checkpoint antibodies. *Ann. Oncol.*, **26** (12), mdv383.

99 Chen, Y.Y., Jensen, M.C., and Smolke, C.D. (2010) Genetic control of mammalian T-cell proliferation with synthetic RNA regulatory systems. *Proc. Natl. Acad. Sci. U.S.A.*, **107** (19), 8531–8536.

100 Di Stasi, A., Tey, S.-K., Dotti, G. *et al.* (2011) Inducible apoptosis as a safety switch for adoptive cell therapy. *N. Engl. J. Med.*, **365** (18), 1673–1683.

101 Wilson, D.S. and Szostak, J.W. (1999) In vitro selection of functional nucleic acids. *Annu. Rev. Biochem.*, **68** (1), 611–647.

102 Bonini, C., Ferrari, G., Verzeletti, S. *et al.* (1997) HSV-TK gene transfer into donor lymphocytes for control of allogeneic graft-versus-leukemia. *Science*, **276** (5319), 1719–1724.

103 Miller, W., Flynn, P., McCullough, J. *et al.* (1986) Cytomegalovirus infection after bone marrow transplantation: an association with acute graft-v-host disease. *Blood*, **67** (4), 1162–1167.

104 Straathof, K.C., Spencer, D.M., Sutton, R.E., and Rooney, C.M. (2003) Suicide genes as safety switches in T lymphocytes. *Cytotherapy*, **5** (3), 227–230.

105 Straathof, K.C., Pulè, M.A., Yotnda, P. *et al.* (2005) An inducible caspase 9 safety switch for T-cell therapy. *Blood*, **105** (11), 4247–4254.

106 Hoyos, V., Savoldo, B., Quintarelli, C. *et al.* (2010) Engineering CD19-specific T lymphocytes with interleukin-15 and a suicide gene to enhance their anti-lymphoma/leukemia effects and safety. *Leukemia*, **24** (6), 1160–1170.

107 Porter, D.L., Levine, B.L., Kalos, M. *et al.* (2011) Chimeric antigen receptor-modified T cells in chronic lymphoid leukemia. *N. Engl. J. Med.*, **365** (8), 725–733.

108 Morgan, R.A., Yang, J.C., Kitano, M. *et al.* (2010) Case report of a serious adverse event following the administration of T cells transduced with a chimeric antigen receptor recognizing ERBB2. *Mol. Ther.*, **18** (4), 843–851.

109 Kalos, M., Levine, B.L., Porter, D.L. *et al.* (2011) T cells with chimeric antigen receptors have potent antitumor effects and can establish memory in patients with advanced leukemia. *Sci. Transl. Med.*, **3** (95), 95ra73.

110 Wei, P., Wong, W.W., Park, J.S. *et al.* (2012) Bacterial virulence proteins as tools to rewire kinase pathways in yeast and immune cells. *Naturr*, **488** (7411), 384–388.

111 Rosenberg, S.A. (2014) Finding suitable targets is the major obstacle to cancer gene therapy. *Cancer Gene Ther.*, **21** (2), 45–47.

112 Hinrichs, C.S. and Restifo, N.P. (2013) Reassessing target antigens for adoptive T-cell therapy. *Nat. Biotechnol.*, **31** (11), 999–1008.

113 Ho, P., Ede, C., and Chen, Y.Y. (2017) Modularly constructed synthetic granzyme B molecule enables interrogation of intracellular proteases for targeted cytotoxicity. *ACS Synth. Biol.*, **6**, 1484–1495.

114 Hasty, J., Dolnik, M., Rottschäfer, V., and Collins, J.J. (2002) Synthetic gene network for entraining and amplifying cellular oscillations. *Phys. Rev. Lett.*, **88** (14), 148101.

115 Elowitz, M.B. and Leibler, S. (2000) A synthetic oscillatory network of transcriptional regulators. *Nature*, **403** (6767), 335–338.

116 You, L., Cox, R.S., Weiss, R., and Arnold, F.H. (2004) Programmed population control by cell–cell communication and regulated killing. *Nature*, **428** (6985), 868–871.

117 Ajo-Franklin, C.M., Drubin, D.A., Eskin, J.A. *et al.* (2007) Rational design of memory in eukaryotic cells. *Genes Dev.*, **21** (18), 2271–2276.

118 Qasim, W., Zhan, H., Samarasinghe, S. *et al.* (2017) Molecular remission of infant B-ALL after infusion of universal TALEN gene-edited CAR T cells. *Sci. Transl. Med.*, **9** (374), eaaj2013.

119 Wei, K.Y., Chen, Y.Y., and Smolke, C.D. (2013) A yeast-based rapid prototype platform for gene control elements in mammalian cells. *Biotechnol. Bioeng.*, **110** (4), 1201–1210.

120 Ede, C., Chen, X., Lin, M.-Y., and Chen, Y.Y. (2016) Quantitative analyses of core promoters enable precise engineering of regulated gene expression in mammalian cells. *ACS Synth. Biol.*, **5** (5), 395–404.

121 Eyquem, J., Mansilla-Soto, J., Giavridis, T. *et al.* (2017) Targeting a CAR to the TRAC locus with CRISPR/Cas9 enhances tumour rejection. *Nature*, **543** (7643), 113–117.

122 MacLeod, D.T., Antony, J., Martin, A.J. *et al.* (2017) Integration of a CD19 CAR into the TCR alpha chain locus streamlines production of allogeneic gene-edited CAR T cells. *Mol. Ther.*, **25** (4), 949–961.

123 Xie, Z., Wroblewska, L., Prochazka, L. *et al.* (2011) Multi-input RNAi-based logic circuit for identification of specific cancer cells. *Science*, **333** (6047), 1307–1311.

Part V

Societal Ramifications of Synthetic Biology

18

Synthetic Biology: From Genetic Engineering 2.0 to Responsible Research and Innovation

Lei Pei and Markus Schmidt

Biofaction KG, Technology Assessment, Kundmanngasse 39/12, Wien, 1030, Austria

18.1 Introduction

This chapter summarizes recent developments in the field of public perception and public engagement and the attempt to apply the concept of "responsible research and innovation" (RRI) to synthetic biology (SB). Albeit the term synthetic biology – in its contemporary version – has been around for about a decade, the field itself can be considered as a continuous development of genetic engineering (GE), a research field established in 1970s [1], although the term synthetic biology itself was coined already in 1910 by French scientist Stephane Leduc. GE is defined as "the intentional manipulation of an organism's genetic material using tools that cut, move, and reattach (recombine) DNA segments within and across different organisms" [1]. SB is developed based on the experience and knowledge of GE [2], yet tools and approaches of SB differ from those of GE as SB attempts to build more sophisticated biological systems [3]. Thus, SB can be seen as the second edition of GE, GE 2.0, as a "new way to organize and construct the art of genetic engineering" that "enforces the traditional engineering concepts of modularity and standardization and adapts logical operator structures from information processing" [4]. Since the early onset of this technology, the GE has faced a lot of skeptics from different stakeholders, including the research community itself, nongovernmental organizations (NGOs), and regulatory bodies. In the early GE development, the public as well as scientists shared similar concerns on how to conduct the research. Along with the growth and development of GE, oversight efforts have been developed to address these concerns at least since 1975 [5].

Although SB uses recombinant DNA techniques to engineer genetic circuits, parts, devices, and the whole systems, it differs from GE. In GE, the principle approach is more likely a "copy and paste" of the naturally existed traits from "donor" to "recipient." Yet as a GE 2.0 version, SB is enabling scientists and engineers more freedom to "compose" contents based on design.

This would entail a deeper metabolic engineering [6], the definition of a minimal genome [7–9], the construction of protocells [10, 11], and the creation of

Synthetic Biology: Parts, Devices and Applications, First Edition. Edited by Christina Smolke.

noncanonical biochemistries [12–17]. SB is an interdisciplinary research field, involving scientists across both science and engineering [4], while GE is known to be a discipline of life science. Another difference between SB and GE is the consideration of societal concerns at a very early stage of development. Generally speaking, the fact that the world today, with the Internet and social media, civil society groups with diverse needs and concerns, means that we encounter a very different *Zeitgeist* today than in the 1970s. But what does this mean for SB? Will SB be just another reenactment of the GE debate from the 1980s and 1990s, or will the debate be carried out in a totally different way?

In this chapter, the public perception on SB and the societal ramifications of its applications will be reviewed. We will analyze how public perceptions toward SB have developed over the years. Then we will look at the contingencies that frame the debate about the technology with a special emphasis on the comparative science and engineering fields. Last but not least, in order to address some of the concerns raised within the open dialogues on SB, the idea to carry out RRI in SB will be introduced.

18.2 Public Perception of the Nascent Field of Synthetic Biology

According to many scientists and funding agencies, SB is believed to hold great potential for applications in multiple economic areas and thus may have significant ramifications for society.

Learning from the history of GE, especially with regard to genetically modified (GM) crops in Europe, the opinions of the prospective end users and end consumers cannot be ignored in SB. Even the most techno-optimistic engineer realizes that he is not working in a societal vacuum but is part of a societal fabric that relates not only to public funding decisions but also to the way the research is done in the lab. Compared to GM crops or the human genome project, where the technology was developed first and then the implications for society, which were discussed rather "downstream," SB demonstrates a new paradigm where societal issues are placed more "upstream." The idea is that strong concerns and objections would appear on the radar screen early on and "appropriate" measures could be taken to deal with it, instead of fully developing a technology in total ignorance of its societal reaction, having to risk the burial of a whole suite of technologies and wasting millions of taxpayers R&D money (as in the case of GM crops in Europe). What appropriate means in this context is another important aspect, which will be discussed under Section 18.4.

One way to find out what "the public thinks" about SB (or any other new and emerging technology) is by conducting public perception surveys. This could be phone call or face-to-face interviews or written questionnaires (sometimes complemented by focus groups, where about a dozen people have a moderated discussion). The advantage of this approach is the relative ease with which to get some first data. The downside, however, is that the results can only be regarded as a rough momentous observation and a deeper understanding of the rationale behind those perceptions is not always possible.

Although SB is still a rather young field, a number of projects and social science research groups have carried out studies on public perception in Europe and the United States. These investigations have been conducted with different methods, ranging from phone surveys and focus group studies to public dialogues, while the sample sizes of these studies were also varied. It is worth to point out that it is difficult to do quantitative comparisons on these data. What we intend to do here is thus merely a summary of these findings on public perception on SB.

18.2.1 Perception of Synthetic Biology in the United States

From year 2008 to 2010, three consecutive surveys were conducted by the Hart Research Associates; and another one was in 2013 [18]. These surveys provided findings on what were the public perceptions on SB.

Awareness of the technology: The public awareness of SB has increase steadily. Those who heard a lot or some of the technology increased from a bit <1 in 10 earlier to nearly 1 in 4 now (9% in year 2008, 22% in year 2009, 26% in year 2010, and 23% in year 2013). This trend of awareness of the technology might reflect the development in the research field. The promise of SB in harnessing biomass to useful products [6, 19–21] and the creation of the first synthetic cell made it to the headlines of mainstream media [22]. The public exposure to newspaper articles or other media types increased especially in 2010, the year of the Venter Institute publication of the so-called synthetic cell. While in the recent years, as no "thrilling" media news came out of SB, the result was a slightly less marked public awareness.

Imaging the technology: Result from the survey of year 2013 showed that nearly one third (31%) of the public surveyed associated the science with something unnatural, man-made, and artificial. 15% linked the science to generate new life via genetic manipulation. The rest was on possible applications in medical science (10% on prosthetics and 6% for new medicine), agriculture (6%), and basic science (5%). The linkage between SB and something man-made and artificial might be resulted from the term "synthetic," which is traditionally linked to man-made chemicals. It might also be the result from the channel the public learned about the technology, for example, from the media coverage on synthetic cells.

Risk and benefits of the technology: Based on the level of awareness of the technology, those who heard nothing showed higher uncertainty in judging the risk–benefit issue (49% vs 23% of those who heard a little and 18% of those who heard a lot or some). From those who heard about the technology, a lot, some or a little, a majority considered the risks and benefits of equal importance (37% and 40%, respectively). For those who heard a lot or some, the positive thinking (28% of net benefit overweight) was more than the negative thinking (17% of net risk overweight). After providing information about SB, the uncertainty reduced (from 27% down to 5%) from the people in the informed group, yet the negative thinking increased from 15% before informed to 33% post informed. It suggested that the public formed their judgment on a

new technology at least partially based on information they learned. With the limited information (just by short information without further supporting information), the public tended to be more skeptical toward the new comer. This finding somehow runs against the belief often voiced by SB scientists that the more the public knows about the technology, the more they like it. The surveys show that this is not automatically the case.

Oversight of the technology: By comparing the survey result of year 2013 with year 2010, it showed that there is a shift of public opinion on how SB should be regulated. More people considered voluntary research guideline as adequate: 43% of them in year 2013 rising from the 36% in year 2010. A detailed decision on the opinions in year 2013 showed that those who favored government regulation had more confidence in federal government to maximize benefits/minimize risks (59%), while those who favored voluntary guideline showed only 33% confidence on government regulation. Although lack of consensus on how SB should be regulated, the majority of people (two thirds) showed support for SB research instead of placing a ban due to lack of information on risks (one third). This attitude remained the same in the latest two surveys (years 2010 and 2013). More support for the technology to go ahead (88%) came from people who held the opinion that benefits outweighed risks, while those who believed risks outweighing benefits were keen to ban the research (61%). Regarding the most problematic issues, the ranking was as follows: potential to create biological weapons (28%), moral concern to create artificial life (27%), harm to human health (20%), and damage to the environment (12%). An interesting finding from the latest survey was that there was very low awareness of the do-it-yourself biology (DIYBio) movement among the public (only 7%), although this is a grassroots movement supposed to encourage public engagement in research through so-called citizen scientists.

Other studies: Besides the surveys mentioned earlier, the public attitudes toward SB from these surveys were further analyzed [23–25]. Pauwels summarized the two clear findings from the SB surveys [24]. The first is that most people know little or nothing about SB. Second, notwithstanding this lack of knowledge, respondents are likely to venture some remark about what they think SB is and the trade-off between potential benefits and potential risks. This is common for the public perception on other technologies as well due to science literacy. Analogous to cloning, GE and stem cell research were recurrent in the dissemination of SB in the science publications and the public outreach materials. More frames and comparators of SB will be reviewed in Section 18.3. The potential applications seem to be another decisive factor in shifting public perception of SB. Finally, the acceptance of the risk–benefit trade-off of SB seems to depend on an oversight structure that would manage the unknowns, the human and environmental concerns, and their long-term effects. It showed that additional investigations were needed to identify other factors that would shape public perceptions about SB, its potential benefits, and its potential risks. Comparison between the US survey and the UK public dialogue was conducted and that the awareness of SB grew significantly in the United States while the UK dialogue indicated a "conditional support" for SB [26]. The development of public perception on SB was also studied by comparing the trends

in press coverage of SB in the United States to those in parts of Europe [25]. It showed that news stories in the United States mentioned more potential benefits (51%, news coverage from year 2003 to 2008) of SB than potential risks (44%), while the European presses mentioned more risks (59%, from year 2003 to 2007) than potential benefit (28%).

18.2.2 Perception of Synthetic Biology in Europe

18.2.2.1 European Union

The public attitudes toward biotech and the life sciences in Europe have been assessed by the Eurobarometer surveys. The recent Eurobarometer on this topic was conducted in 2010 based on representative samples from 32 European countries [27, 28]. The analysis on the survey showed that the people in Europe were largely unaware of SB – only 17% of those participated in the survey heard of SB – which means a low level of awareness. Regarding GE in general, there were concerns on products, particularly food from the GE technology [28, 29]. Among these concerns, there were common perceptions that the GE food was probably unsafe or even harmful; there was also concern on safety due to possible horizontal gene transfer. However, the public attitudes toward novel technology were not totally negative. The survey showed the public believed that research on biofuels (an application developed by SB approaches) should be supported. The primary concern on SB was the information about the possible risk (63%). A majority of the public would also like to know more about the claimed benefit (52%). Other concerns were who would benefit and who would bear the risks (40%), scientific progress in the field (31%), regulation (29%), funding (24%), and societal issues (16%) [28]. Due to the unawareness of the technology and more enthusiasm for the novel field, the public considered the regulation of SB should be science based (left for the scientific experts) but with the necessary oversight from the authority; however, when ethics and social values were involved, the public involvement should be included in decision-making [29]. When asking how SB should be regulated, more than half preferred scientific evidence (52%) over moral or ethical issues (34%). And the public preferred more expert advices for the decision about SB (59%) than what the majority (lay people) would think (29%). A majority (77%) agreed that SB should be tightly regulated by the government [28]. Within 2014–2015, the expert committees from European Commission (EC) conducted three public consultations related to SB, covering issues on the definition on SB, risk assessment methods and safety aspects, and SB-related risks to the environment and biodiversity and research priorities [30–32]. The reports from these public consultations, although the opinions were most likely from closely related stakeholders in the field, paved way for further dissemination of SB in Europe.

18.2.2.2 Austria

A study on communicating SB from scientists via the media to the public was conducted by the Austrian COSY (Communicating Synthetic Biology) project in 2008 [33]. Press releases written by the scientists on their work were reviewed by four journalists from major Austrian newspapers and magazines. The journalists

then wrote articles based on these materials, which were used as topics to be discussed by eight focus groups with member from the Austrian public. This study showed two important observations of science communication from the scientists via the media to the general public. The first observation indicated that journalists focused on and selected more for real-world applications of SB (pharmaceuticals, biofuels, etc.) than the abstract key scientific and engineering concepts (such as standardization, modularization, etc.). As a result the very key aspects that distinguish SB from GE were not disseminated properly to the public via the social media (by the journalists, in this case), and thus the laypeople could not see any difference between GE and SB, believing that SB was just another name for GE. The second observation concerned the relation between information and attitude. Before the participants got to read the articles, their opinion toward SB was neutral, neither positive nor negative. But after reading the articles, and partly due to the link made to GE, two groups became very negative, and two groups became very positive toward SB, whereas four groups still had a rather neutral or uninterested opinion about SB (see Figure 18.1). So the assumption that "the more they know, they more they like it" cannot be established based on these empirical results. It turned out that the attitude toward SB would likely be polarized after more information was provided. The information in the focus groups was not taken in a neutral way, but rather the social identity of the individual would influence the revision of the early attitude. This might explain the polarization of attitudes toward SB in the Austrian laypeople.

A media analysis on SB was done on the German-language media articles published between 2004 and 2008 [35]. It showed that the media reported about SB focused more on the positive potential and less on the risks. The definition of SB was introduced to the public along with possible applications. Journalists used common metaphors to define SB such as "biological engineer," "playing Lego," or "redesign of life," while the common phrases were related to the terms "machines," "factories," "computer engineering," and "creation." The analysis from this study

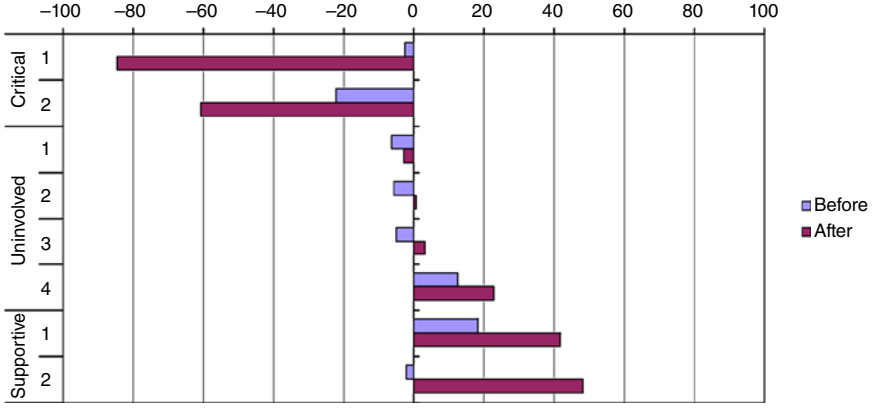


Figure 18.1 Focus group's evaluation of SB, before and after they receive information. The x-axis means: −100 totally opposed; 0 neutral; +100 totally endorsing. Consequences of media information uptake and deliberation: focus groups' symbolic coping with synthetic biology [34].

showed again the way of science communication of the media on an emerging technology, which might not be able to disseminate the technology properly to the public due to lack of information on the key scientific issues while preferring common metaphors.

A rather unusual way of trying to understand how lay people would react to SB was done in relation to one of the first SB art exhibitions. The exhibition "synth-ethic" showed 10 artworks from 10 international bio-artists in Vienna in May and June 2011 [36]. The artwork dealt with various aspects of SB, the use of biobricks not attempted by engineers, creation of protocells, modifying ecological networks, potential environmental release, the meaning of synthesis as opposed to analysis, etc (see http://www.biofaction.com/synth-ethic/). During the exhibition gallery visitors were interviewed about their perception of the artwork and the relation between art, science, and society. The results showed that people had little ethical problems with the artwork as long as it entailed the use and modification of lower life forms (bacteria, plants), but they were more concerned as the bio-objects moved up the evolutionary ladder toward birds, mammals, and even humans. An innate key concept seems to be the need to be able to keep different categories separated from each other, and any crossing of boundaries triggered uneasiness. Boundaries could be crossed in an ethical sense by modifying and designing mammals and humans or by crossing two different living entities (hybrids) or by crossing organisms and machines [37]. Any attempt to cross well-established boundaries of lay people's naïve view on biology could result in public resistance.

18.2.2.3 Germany

Three German research organizations – the German Research Foundation (DFG), the German Academy of Science and Engineering (acatech), and the German National Academy of Sciences (Leopoldina) – published a position article to outline their strategies to SB while suggesting a broadly based scientific and public debate on SB [38]. It suggested that SB would make major contributions to the society while bearing risks, such as legal aspects, biosafety and biosecurity, commercial use, and ethics. While German scientists are active in the research field [39], the German public, similar to its smaller neighbor Austria, holds skeptic views on GMOs. It was feared that crossing "the boundaries between living matter and technically constructed matter" would cause public concern and that ethical boundaries were broken down as well. In their role as funding agencies in Germany, they proposed that the activities supported by public funding should "guarantee transparency by means of communication that will foster public acceptance of this research field." The ethical issues would need to be debated by the public further based on the seven hypotheses and goals [38]:

1. The definition of life.
2. The factors that determine the preconceived understanding of life.
3. The description of entities.
4. Moral arguments on applications of SB taking into consideration basic rights.
5. Fundamental ethical objects against the applications of SB.

6. The debate on self-regulation of science.
7. All the discussions would be on "a comprehensive, interdisciplinary and intercontextual" basis.

18.2.2.4 Netherlands

The societal issues around SB were studied by the Netherlands Commission on Genetic Modification (COGEM). Their report published in 2010 analyzed the developments in SB and intended to answer the questions on when and how governments would have to anticipate the public debate on SB in order to prepare the future developments in the field [40]. In the Netherlands, the emergence of SB was subject to a public debate, which seemingly reiterated the old debates on biotechnology.

High controversies are always raised for new technologies with high expectations, where SB is no exception. It was believed that there was a gap between the technical expectations and the reality, as no concrete SB applications had yet reached the market. A problem identified was that at a time when the hype was dominated and reported by media, little information about specific societal implications was available, and later on when this information was available, the topic disappeared from the media and the public debate.

The COGEM report concluded that thus "technology assessment needs to facilitate the societal-ethical debate when media attention, and thus the visibility of the technological developments, declines." It brought up a situation of how the public debates on SB should be conducted: the scientists speculated in the media about future developments in SB, and the media played "host to an exchange of 'dream' and 'doom' scenarios." It suggested that the gap between available information and hype-based media attention should be closed by using technology assessment to facilitate public debate. For example, a technology assessment was done on SB, pointing out new dimensions to old questions in public debates [41]. The issues identified were biosafety, misuse/bioterrorism, intellectual property, and ethics, in comparison between GM and SB. The challenges that SB raised were new questions and uncertainties about risks, difficulties in monitoring misuse and research on potentially harmful organisms, new hurdles for research and innovations, and blurring boundary between life and machines. These issues and challenges should be primary topics for the public debates. Meanwhile, different technology/policy processes should be used in different stages of societal–ethical discussions. At the early stage, the public debate would be initiated through the expectations articulated by the scientists. The introduction (mostly promises and expectations) of the emerging technology prompted the general public to form a perception. The real developments in the field – breakthrough or failures – would prompt the public to revise what they were first told. During the growing stage of a technology, it was the achievement in the field that led to public debates. While the concrete application was absent, it was not easy for the government to address the possible issues in advance and steer developments accordingly. The public debates on this stage should have clear goals, with objectives either to steer the direction of the technological development or to gauge public support for the development or as an input to shape/support policy.

18.2.2.5 United Kingdom

In the United Kingdom, public perceptions on SB were studied by the Royal Academy of Engineering (RAE) and the Biotechnology and Biological Sciences Research Council (BBSRC) with input from the Engineering and Physical Sciences Research Council (EPSRC) in 2009. The report from RAE was based on a dialogue activity with 16 members of the public and a nationwide representative survey of 1000 adults aged 18 and over [42]. The perceptions of laypeople about the scientific research, awareness, and understandings of SB were investigated. The report showed that in the United Kingdom the awareness of SB was low. This is similar to the findings in the United States and the rest of Europe – nearly two thirds had never heard of SB, and for the one third who heard of SB, only 10% among this one third heard a lot (or which is 3% of the all answers of the survey), 57% a little (or 19% in total), and 33% only the term (or 10% in total). The words linked to SB were "artificial," "unnatural," and "man-made." While studying the public attitudes toward creating, modifying life, and totally man-made organisms, the majority of the respondents were positive about creating microorganisms to produce medicines and biofuels, which were also found in the Eurobarometer. Regarding issues around SB, the survey showed that there were biosafety concerns on SB applications involving environmental release, while there was comparatively little concern about the biosecurity issues SB might bring with.

The BBSRC and EPSRC published a report outlining the most important findings around the Synthetic Biology Public Dialogue [43]. This dialogue was conducted by TNS-BMRB with 41 stakeholder interviews, involved 160 members of the public and specialists on science and governance. In the United Kingdom it is highly expected that SB could address some challenges for the whole society. Yet how to foster such a science should take into account the social context. The dialogue was conducted among the interested groups from the public, people from the research community and other stakeholders, to explore the public expectations, concerns, and aspirations around SB. The major findings from this dialogue were as follows: people were both excited and scared by the potential of SB; they were concerned about adequate regulations and preferred international regulations on SB, particularly for those applications that (might) affect the environment; and the public was concerned about the motivation of scientists who were asked to consider the wider impacts of their work. The UK dialogue revealed the important role of the public debate on SB and showed its impacts in dissemination, awareness of the issues raised in the dialogue, and the needs for public engagement.

The UK dialogue also showed the different views on SB from different stakeholders. For example, the researchers from the academic field tended to "rebrand" their research with SB to attract funding, while the researchers from the industry tended to avoid the SB label due to the negative perception of "synthetic" among the lay public. The social scientists, NGOs, and the consumer groups viewed that the development of SB was driven by the interest from the large corporations [44]. However, these different views did not hinder all the stakeholders to agree on the value of public engagement. A dialogue engaging all the stakeholders will

provide communication channels to build the responsible research in SB, which we will review in Section 18.4.

18.2.3 Opinions from Concerned Civil Society Groups

Attentions on SB are not limited within the academic community (social science, science and technology studies, technology assessment, etc.) and regulatory bodies. Concerned groups, especially environmental NGOs [45], for example, the Action Group on Erosion, Technology and Concentration (ETC), have conducted a couple of studies on SB and GE since 2006 [46–49]. In 2006, the ETC Group and other NGOs published an open letter calling for a societal debate on socioeconomic, security, health, environmental, and human rights implications of SB [50]. In one of their reports, they argued that the advocators of SB intended to "avoid public scrutiny by asserting that it is impossible to clearly distinguish their work from earlier advances in recombinant DNA technology (genetic engineering)" [47]. They recommended that public dialogue should be encouraged and the potential risks should be made transparent. While promoting SB could contribute to "the green economy," they argued that "a full global public debate on all of the socioeconomic, environmental and ethical issues related to biomass use, synthetic biology, and the governance of new and emerging technologies in general" was needed [51]. ETC together with other NGOs such as Friends of the Earth U.S. and International Center for Technology Assessment (ICTA) published the suggested principles for the oversight of SB. According to them, "full public participation at every level" should be included in the oversight of SB, and "full disclosure to the public of the nature of the synthetic organism" should be a prerequisite for commercialization or environmental releasing of any SB product (ETC et al. 2012). As recently as October 2012, ETC together with Friends of the Earth managed to get their concerns heard at the COP 11 UN meeting on the Convention on Biological Diversity (CBD) in Hyderabad, India. With 193 nation states represented, the representative of the Philippines asked for a moratorium on SB (initially suggested by the NGOs), which was then rejected by the other states. As a response to the critical views a final statement by all nations asking for a cautious approach to SB followed. The opinions from the concerned groups show the needs not only for public engagement but also for open access to the technology. These issues are key to the framework of RRI, which will be reviewed in Section 18.4.

18.3 Frames and Comparators

As we have shown in Section 18.2, comparisons between SB and GE are widely used when scientists communicate with their peers and with the public. Thus SB can be seen as GE 2.0. There are, however, strong indications that SB – in science and in the public debate – goes beyond a mere continuation of GE. Such debates are subject to dominant frames, because otherwise it would not be possible to discuss anything [52, 53]. For a development of a debate, it is necessary to develop a common understanding of what is to be considered relevant and which form of

argumentation is deemed legitimate. Without such an understanding, a debate is dysfunctional, and potential participants are unable to have a discussion in the first place [54, 55]. They do need to identify a common frame, under which the debate can be held [56].

The choice of a dominant frame does not determine the fate of a debate. But it has implications for the selection of relevant expertise, of the kind of stakeholders to be invited, of the type of measures to be taken, etc. For example, the debate about green biotechnology in Europe was mostly held under a risk frame, that is, arguments about risk for human health and the environment were deemed more relevant than economic equity or ethical concerns. Consequently, scientists were asked about the probability of risks, and prior risk assessment was made mandatory. In the stem cell debate, an ethics frame prevailed, and arguments over the sanctity of embryonic life were considered more important than health risks. The expertise taken on board in the negotiations included those of ethicists and clergymen, and measures included a ban on some forms of research. Yet another frame different from risk and ethics is the economic frame, emphasizing the opportunities for future benefits, growth, and opportunities for the economy.

In principle, other frames might be conceivable. However empirically, in technology debates they are most frequent: media analyses of technology controversies revealed "basic frames" that are not fundamentally different ones [57].

For an upstream debate on emerging technologies such as SB, dominant frames do not readily emerge from the issue itself, as this issue still is vague in its properties and consequences. Analogies to other technologies having left a mark in the publics' imagination come in handy here. The frames of the past debate on the older comparator technology influence those developing in the debate over the new technology. In practice, frames are often "copied" from a comparator debate and "pasted" into the new one: dominant arguments and the choice of issues relevant in the debate over the older technology serve as a blueprint for debating the implications of the new technology [56]. We might as well call it a "recombinant debate."

Many observers have expressed the assumption that SB would follow the same development as GE in the 1980s and 1990s, hence the word creation GE 2.0. SB, however, as a true interdisciplinary and converging technology has been linked not only to biotechnology but also to nanotechnology and information technology (IT) [56, 58]. Each comparator conveys different aspects, expectations, hopes, and fears; and the dominant debates are held under partly or entirely different frames, respectively. Each comparator entails a unique way to understand and interpret the technology at stake. For biotechnology, the comparator stands for "technology as conflict"; in the case of nanotechnology, it is "technology as progress"; and for IT it is "technology as gadget." The terms "conflict," "progress," and "gadget" are used here only to catch the main meaning of the frame in single term (see Figure 18.2). Thus, "If a comparator becomes dominant, i.e. obvious to many experts, stakeholders and members of the public it might influence the course of a debate 'out there' through suggesting one or more dominant frames. They will reflect the encompassing nature of the debate through their implicit conceptualization of the public: 'technology as conflict' goes along with the

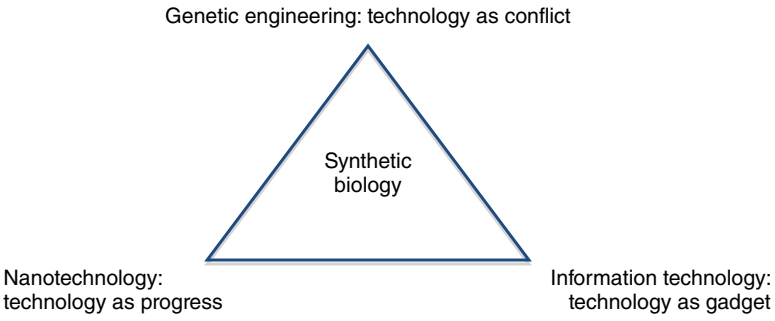


Figure 18.2 The dominant comparator for SB could come from either of three preexisting technology debates.

public to be taken seriously; seen through the glasses of 'technology as progress', the public appears as an entity to be mastered through appropriate means; and with 'technology as gadget' the public is seen as a player in the technology's own team, so to say" [56].

18.3.1 Genetic Engineering: Technology as Conflict

GM crops or "green" biotechnology have been subject to adverse public perception in some countries. It is therefore not surprising that critical NGOs refer to GE as a comparator for SB [47], painting a dark picture with SB being internalized into the agenda of big business to exploit natural resources even more aggressively. The ETC Group dubbed SB to be "extreme genetic engineering" and underline the risks and inherent conflicts, such as intellectual property rights (IPR), economic and power concentrations, environmental safety, and rural livelihoods. As a general rule, those environmental NGOs having addressed SB so far tended to extrapolate arguments against various forms of biotechnology to future applications of SB (ETC et al. 2012).

Policy refers to the GE comparator mostly in the form of a menace: "the same" as with GM food (i.e., a failed implementation due to public rejection) must be prevented. The IRGC report of 2010 (p. 37) described this reaction common among experts and policy makers as "… the 'fear of the fear of the public' – a concern among those working on synthetic biology that the kind of public response to GM crops that emerged in Europe in the late 1990s would be transferred, perhaps in a more virulent form, to synthetic biology." The problem, accordingly, lies in how to "… find ways of reconciling fundamentally conflicting values or ideologies." "… there are strong differences of opinion at the outset of a debate, it is hard to manage the process in such a way as to avoid further polarization of views and exacerbation of conflict" – exactly as with GM food [59].

Communication strategies by many scientists and those from the industry are to try to emphasize the difference to conventional biotechnology/GE. This may be related to presenting a promising new field to funding agencies. On the other hand, it has been pointed out that SB is an extension to GE transgressing past approaches but proceeding on the same avenue toward artificialness. In their

approach to the public, allusions to the biotechnology conflict in Europe can be found, although many prominent scientists come from the United States where biotechnology has not met with particular problems among the public. In Europe, in contrast to fears among policy makers, SB has not met with strong objections so far. A reason might be that it has not impinged on food, and food issues are used to be major conflict triggers not only regarding GE.

18.3.2 Nanotechnology: Technology as Progress

Nanotechnology is an emerging technology par excellence, bearing high expectations and benefiting from massive public funding – the EC alone, for example, spent €3.5 billion through the 7th Framework Programme (FP7). Regarding potential risk, both allegations and serious concerns have been addressed more professionally than with biotechnology in its early days. Assessments mostly resulted in identifying far-reaching knowledge gaps to be filled in incrementally but rapidly. In contrast to the perception of some technical experts and policy makers, press coverage has not particularly focused on risk so far; rather, the potentials for huge benefits have been mostly to the fore [60]. Despite many speculations that nanotechnology might elicit concerns similar to GMOs (and occasional demonstrations limited mainly to France), it succeeded to evade the public rejection trap.

To address some negative speculations on nanotechnology, a variety of public engagement exercises have been set up (see e.g., [61]). Apart from more academic social science research, information initiatives such as the "nanoTruck" in Germany, science fairs, and similar upstream outreach activities as well as a number of participatory events of different forms are belonging to a new way of successfully introducing a novel technology "in a responsible way." Among other outcomes, this focus helped coin the term "responsible research and innovation" the EC subscribed to also for other technological areas [62], which will be further discussed in the next section.

18.3.3 Information Technology: Technology as Gadget

IT or computer technology changed our life over the last decades in an unprecedented way. Few technologies had a similar impact on modern society. Computers govern virtually every aspect of our modern existence and cause an explosion in productivity. Initial resentments were overcome quickly, and IT has developed into a synonym for the most powerful, pervasive, and, at the same time, "cool" technology imaginable. Gadgets and toys galore have contributed to this image, and possessing the newest product has become the most relevant status symbol. There is a critical debate on the aspects such as intellectual property, privacy, or cybercrime, to name but a few, yet the technology as such is established beyond any question.

SB can be considered as an IT too, only using a different medium, namely, DNA base sequences rather than software codes. Protagonists stress the IT analogy to a remarkable extent, and many pertinent examples and apparent similarities between SB and IT appear in the literature. The analogies mostly refer to

elements of the technologies themselves even if they derive from entirely different disciplines such as electronic engineering.

The closest link of SB and IT is established through the scientists and engineers involved – many of the original protagonists in SB come from the IT sector. As part of their professional world view, they frequently allude to IT construction elements such as integrated circuits, devices and systems, etc., when talking about biological entities such as genes, biological pathways, cells, and organisms. In addition, they decidedly set out to apply engineering principles in biology, which is also the most frequently used definition for SB. Even the formation of amateur biologists or DIYBio groups comes from a hacker tradition seen in the IT world.

Using the IT frame as a dominant guide for assessing and debating the ramifications of SB, the result – to a great extent – is a predominantly positive, cool, and gadget-like perception of SB. Yet it also calls for addressing safety and security concerns as well as intellectual property issues as those of IT. Thus, fostering responsibility in SB research should also be established alongside the development of the technology, which will be discussed in the following section.

18.3.4 SB: Which Debate to Come?

Since SB is still largely unknown by large parts of the public and contemporary debates are held mainly among experts, it is hard to tell which way the SB debate is going to play out. Will it develop along the lines of the old GE debate, as many environmental NGOs link it to? Or will the nanotechnology or IT comparator frame the debates? We are not aware of any hard facts to determine the future debate about SB. In the light of absence of such hard facts, some scholars investigated artistic expression as a sense of possibility.

A kind of sneak preview of the debate to come was presented by a study of independent SB short films [63]. The authors analyzed (semi-) fictional short films about SB that were shown during the Science, Art and Film Festival BIO·FICTION (see www.bio-fiction.com/videos). In this festival, filmmakers presented their visions of how SB would be taken up by society and their views through the short films. Since artists can to some extent be regarded as cultural psychologists, the depiction of SB in these science fiction/documentary films might as well help us to grasp the first hints of an SB debate to come. Going through the 52 short films from BIO·FICTION, the authors used the input to elaborate an analysis that comes to the conclusion that "representations of SB in the Bio:fiction films confirm with our hypothesis that the debate about SB is not seen as a straight continuation of the debate in biotechnology/genetic engineering. Instead, alternative narrative attractors seem to be dominant. Although we were not able to make a clear case for either technology as progress or technology as gadget, since both aspects played out more or less equally, we could clearly reject the technology as conflict frame [63]."

Analyzing the three main comparators of SB, it shows that SB goes beyond GE 2.0, as indicating from the scientific/technological stance and the early indications of public debates. To facilitate the development of SB and to leash the full

potential of SB, it calls for building a new framework for research – a framework for RRI.

18.4 Toward Responsible Research and Innovation (RRI) in Synthetic Biology

Implementing the RRI approach into SB can help to address the societal needs and challenges of the emerging technology. In the roadmap for SB in the United Kingdom, continuing RRI has been brought up as one of five core themes to achieve a successful outcome of SB in the United Kingdom [64].

RRI has been defined by the EC as "the comprehensive approach of proceeding in research and innovation in ways that allow all stakeholders that are involved in the processes of research and innovation at an early stage (A) to obtain relevant knowledge on the consequences of the outcomes of their actions and on the range of options open to them and (B) to effectively evaluate both outcomes and options in terms of societal needs and moral values and (C) to use these considerations (under A and B) as functional requirements for design and development of new research, products and services" [65].

Also the RRI approach should "be established as a collective, inclusive and system-wide approach." The RRI is considered as "a key pillar in the strategy of the European Union (EU) to create sustainable, inclusive growth and prosperity and address the societal challenges of Europe and the world" [65]. Its objective is to address "the ethical concerns and societal needs in research and innovation," which can contribute to anchoring research and innovation (in the normative dimension), help to deliver the targets set out in Europe 2020 strategy (substantive dimension), and help to improve research administration (instruction dimension). In 2012, the EC issued a call for action plan for societal challenges. And one of the special challenges they aimed for was the RRI in SB. It reasoned that although SB held many significant promises for the society, the public was not yet much aware of this nascent field and the associated regulatory challenges. Thus, "it is essential to establish open dialogue between stakeholders, to understand public concerns and ensure collaborative shaping of the field, aligned with societal needs and expectations" [66]. A dedicated project on RRI in SB has been funded by FP7 to establish an open dialogue between stakeholders concerning the potential benefits and risks of SB and to explore the possibilities for its collaborative shaping on the basis of public participation [67].

It is believed that RRI should be practiced continuously, which will help to ensure the awareness of potential issues and keep the regulatory frameworks up to date with progress in the field. SB is a nascent research field and RRI is a relatively new concept. It will, in no doubt, bring both challenges and opportunities to build an RRI framework for SB. The RRI concept has been promoted by the EU via funding schemes to encourage researchers from both natural science and social science to implement the concept into their research projects.

Here, we will review what implications of RRI will bring into the practice of SB; explore the idea of RRI from several different angles, including engagement,

gender equality, science education, open access, ethics, and governance; and discuss how this framework will be constructed.

18.4.1 Engagement of All Societal Actors – Researchers, Industry, Policy Makers, and Civil Society – and Their Joint Participation in the Research and Innovation

Engagement of all societal actors is key for RRI framework, which will help to bridge the gap between the scientific community and society at large. The European Group on Ethics in Science and New Technologies (EGE) published their opinion article on SB [68]. In this report, the philosophical, anthropological, ethical, legal, social, and scientific issues raised by SB were analyzed from the scientific aspects, legal, governance and policy aspects, and ethical aspects. It pointed out particularly the importance of public involvement and science–society dialogue. The European Academies Science Advisory Council also investigated the scientific and governance implications of SB [69]. It, too, pointed out the importance of raising public awareness on the opportunities and challenges of SB among both the scientific community and with the public, as well as continuous public dialogue to ensure that "endeavours in synthetic biology reflect wide public interests and aspirations." Among the six recommendations provided by the Working Group of Experts, one was on societal engagement, emphasizing the proactive approaches the research society had already applied to encourage and inform the public debate based on the accurate information. Dialogue among all the societal actors is a prerequisite to build a framework of RRI. Implementing RRI in SB would provide a unique stage for all societal actors to carry out the dialogues.

A report from the Technology Strategy Board of the United Kingdom outlined the importance of RRI in SB particularly to its transition to industry applications. A responsible innovation framework would require ethical, societal, and regulatory considerations both during the R&D process and during the commercial use. Throughout this process, all the stakeholders would have to get involved [70]. A newly funded project (SYNENERGENE) by the EC under the call for Science in Society will provide some insight how to such a framework to foster the growth of SB. This project will be conducted jointly by 27 partners around Europe, the United States, and Canada, which will bring together a wide range of scientists, regulators, NGOs, companies, and other stakeholders to act together to raise public awareness of SB and to get the stakeholders involved and encourage public discourse and policy in an international context.

As mentioned in the earlier sections, SB will have the potential to bring applications to the society, and people from different background would have different concerns on these applications. Thus, RRI aims to build "transparent, interactive processes in which societal actors and innovators become mutually responsive to each other with a view on the ethical acceptability, sustainability and societal desirability of the innovation process and its marketable products" [62], ideally bringing together societal actors with different interests and values to reach a consistent strategy for developing the technologies and their products.

In RRI the importance to engage all stakeholders, including the public, was emphasized [71]. By engaging all the social actors, it will help to build up public acceptance of innovation, to fulfill the government's responsibility to give citizens opportunity to express their opinions, and to make sure public and civil society stakeholders are also co-players of research and innovation. The broad-spectrum public engagement would make research and innovation more effective. The matured public perception on the technology will be important for future applications SB will develop. The advances of SB might make the knowledge and technology available to the amateur scientists, making them possible co-contributors. How to get the public involved is still a challenge, and the policy makers need to find solutions to make public involvement efficient and to assist the public to form their opinion. Developing proper models of SB to engage the societal actors should learn from the experience obtained from those of GE, nanotechnology, and IT (models of conflict, progress, and gadget). In an opinion paper by Nerlich and Mcleod, they argued that raising awareness on SB should be responsibly, in short, raising awareness of SB by responsible communication while comparing to the case study on climate change, the awareness of which should be advocated responsibly [72].

18.4.2 Gender Equality

Gender equality is the second key issues for building the RRI framework. In the latest report from the EC on structural changes in research institutes, integrating a gender perspective has been considered as one of the key solution to improve research in the EU [73]. Promoting gender equality in all levels contributes to research excellence and efficiency by making full use of a wider talent pool of human resource. The report brought up gender equality strategy (key steps) for actors at the EU, national and regional level, as well as to gatekeepers of scientific excellence and to universities and scientific institutes. For example, the EC should make gender requirements to all funding programs; dedicated programs should be created to promote structural changes in research institutes; EU should set a good model at the worldwide level regarding gender issues; special unit for gender issues should be reestablished; a high-quality leadership development program should be created targeting experts; and researcher mobility measurement should incorporate gender dimension.

Gender issues have already been studied by the SB society. The Sybhel project has studied how SB might influence the philosophical concepts of human health, which also involved gender aspects, and analyzed gender issues related to SB techniques in one of its work packages [74]. The ESRC Genomics Policy and Forum at the University of Edinburgh run a public engagement program on SB. A Democs card game on SB was designed, and playing Democs game was used as a resource to explore the public engagement of SB with lay publics in Scotland. In their report, the gender of the participants was analyzed in the feedback of the game [75]. However, neither these reports provided a comprehensive understanding of the gender issues in SB. Thus dedicated projects are needed to address these issues.

18.4.3 Science Education

The third key to RRI is science education [76]. Creative learning of fresh ideas will help enhance the current education process to ensure all societal actors can get relevant knowledge and tools to participate and make knowledge-based judgment in the process of research and innovation. The educational activity currently being explored are education initiatives that will promote "a culture of responsibility, participative inquiry, nuanced debate -starting in primary or high schools and including governments, scientists, businesses and civil society" [71]. The same report also suggested the roles different stakeholders should play to enhance sciences. Both the governments and research funders should foster in interdisciplinary cooperation and education. The consideration of ethical issues and societal needs should be addressed through education and training. This would prepare the societal actors better to anticipate ethical concerns and to take these concerns into consideration in the future R&D [65].

SB is a nascent research field and well known for its interdisciplinary nature. The novelty of SB calls for better science education – targeting not only the public at large but also the researchers from other disciplines. Meanwhile, activities around SB, such as the International Genetically Engineered Machine (iGEM) competitions and DIYBio movement, have already provided existing platforms for serving the education purposes. The iGEM competition is a worldwide SB annual competition. It initially aimed at undergraduate university students to promote their interest in this nascent field. But it has now expanded to include divisions for high school students and other interested groups outside the university setting [77]. DIYBio is a growing movement among amateur biologists. They are individuals, or small groups, who conduct biological research outside the conventional institutional setting (such as in academic or industrial facilities) with limited resources. Amateur biologists have little or no formal training in biology [78–82]. Both the iGEM and DIYBio movement will open up new education channels to the public. Both of them call for more supports from the professional society and regulatory bodies to ensure the activities are conducted in efficient and beneficial way [83–88].

Among the many different ways of science education to engage nonscientists in science and technology issues, the use of science games comes handy. For example, BioFaction, as part of the European Science Foundation project on synthetic lantibiotics called "SYNMOD," developed a mobile app game to present the concept and aim of the project in an entertaining and accessible way (see Figure 18.3).

18.4.4 Open Access

SB is a fast-growing field that can be assigned broadly to the knowledge-based bioeconomy. Although SB is still in a nascent stage, the issues on open access have already raised concerns for potential future applications. A study was done to analyze the comparative benefits and pitfalls of open access and patenting issues [89]. As mentioned earlier in the chapter about frames and comparators, SB is also influenced by the IT sector. So it comes as no surprise that some ideas

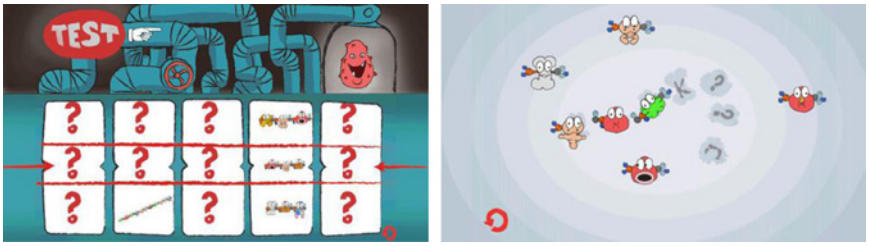


Figure 18.3 The SYNMOD game app allows players to create and combine different peptide modules to design new antibiotics. The game is freely available for iOS X and Android devices. See http://www.biofaction.com/project/synmod-mobile-game/

and practices in SB are influenced by the IT world. One specific example is the uptake and translation of the open access/open source software to the world of biotechnology, a field where so far restrictive IPR have been used [90]. Unlike restricting access to crucial information, some synthetic biologists want to develop an open access to share the information they obtain. That is the theory behind the BioBricks Foundation (BBF), the International Open Facility Advancing Biotechnology (BioFab), the Biological Innovation for an Open Society (BIOS), and the Synthetic Biology Open Language (SBOL).

BBF was founded by the scientists involved with the Registry of Standard Biological Parts, aiming to provide a platform to "ensure that the engineering of biology is conducted in an open and ethical manner to benefit all people and the planet." The Registry of Standard Biological Parts aims to allow interested actors to contribute and access standard genetic components, so-called parts and devices. Recently the BBF published the custom made BioBrick™ Public Agreement, which tries to set up a legal way to ensure open access [91].

BioFab was funded by the National Academy of Sciences (United States) to support an open technology platform and to provide free genetic constructs that can be customized for specific applications by academia and industry.

BIOS was created to "enhance the transparency, accessibility and capability to use all the tools of science, whether patented, open access or public domain." It is believed that the "open access to research" concept will not only increase the transparency in research but also promote free exchange of information. According to its proponents, such transparency will promote development by sharing knowledge among the research community and will help to reduce the misuse of the technology [89].

SBOL is an open source movement for *in silico* representation of genetic designs. SBOL is designed to allow electronic-like exchange designs, to send and retrieve genetic designs to and from the research centers, to facilitate storage of genetic designs, and to embed genetic designs in publications [92]. More and more bioparts have now been registered in the database. A registry software, the Joint BioEnergy Institute Inventory of Composable Elements (JBEI-ICEs), was created to provide a platform to manage the growing information on bioparts. The JBEI-ICE is built to support for distributed interconnected use and to provide well-developed parts storage functionality for other SB software projects.

The open access approach demonstrates not only the willingness of the free flow of information among parts of the scientific community but also the demands from the public to secure the common benefit from the public-funded research. Thus making open access a reality is an important aspect to build the framework of RRI. The challenges for open access are basically twofold: firstly whether it will be sustainable and successfully picked up by "users" and secondly the legal issues of some open source content that (might) overlap with the existing patents.

18.4.5 Ethics

Another key component for the RRI framework is ethics. The shared values among the European society call for RRI to be built respecting fundamental rights and the highest ethical standards [76]. As early as 2006, to stimulate the develop of SB in Europe, the EC funded 18 SB projects through NEST Pathfinder, aiming to stimulate advancements in science as well as to address ethical and safety concerns [93]. Among these projects, SYNBIOSAFE was particularly dedicated to study safety, ethical, and governance issues [94]. A number of other SB ethics-related projects funded by the EC came followed by SYNBIOSAFE later on. The EGE published their opinion article on SB [68]. Ethical issues raised by SB were analyzed by the EGE, including biosafety, biosecurity, justice, and intellectual property issues [68]. Twenty-six recommendations were proposed by the EGE in their opinion article regarding safety (environmental applications, sustainable energy, and healthcare products), security, governance, intellectual property (patent and justice), science and society aspects, and basic research.

A recent study from EC pointed out that there were gaps between research and innovation systems and RRI regarding ethics. The research system failed to consider the ethical and societal aspect sufficiently, and the innovation system often failed to anticipate future societal needs. For both systems, the researchers were often less aware of the ethical and societal impacts of their research activities [65]. To integrate the ethical dimension into the research projects, the EU has asked the researchers to address the ethical questions and questions of social needs (if any) associated with their project in their grant applications and research projects. To further integrate research responsibility into the research projects, the expert group brought up an improved option other than the "business as usual" option: more research funding should be allocated (€79 billion for Horizon 2020 and €2.5 billion for COSME); and the researcher should reflect both ethics and responsibility in their proposals [65]. This option will require RRI to turn into the mainstream of the EU funding programs. The share of trans-/interdisciplinary research should be increased. Furthermore, a special funding should be set up dedicated to RRI research. The importance of ethical consideration in research has been also emphasized a UK study [64]. The funder BBSRC has placed a number of checks and balances to ensure the awareness of the ethical and social issues raised by the funded projects. Examples for the checks of ethical issues are ethical considerations on using animals in an experiment and the potential for misuse/dual use of the knowledge obtained from the projects.

It is believed that guidance on responsible ethical assessment is needed to be vigilant about the harms of an emerging technology and prepared to revise the policy while necessary. This calls for a broad-based ethical framework for SB. A couple of key ethical principles relevant to the social implications of SB should be taken into consideration to evaluate SB and its potential risks and benefits, such as public beneficence, responsible stewardship, intellectual freedom and responsibility, democratic deliberation, and justice and fairness [95]. To apply this broad-based ethical framework to SB, public dialogue on ethical issues of SB is one of the key components. The model developed based on the frames and comparators will be applied to these public dialogue events to provide the participants accurate yet understandable information about the topics.

18.4.6 Governance

Harmonious models for RRI integrating public engagement, gender equality, science education, open access, and ethics can be built with proper governance. The policy makers are the ones who should take action [76]. To clarify the role of authority in regulation of SB, the European Academies Science Advisory Council investigated the scientific and governance implications of SB [69]. It is still in debate if specific policy for SB is needed to advance the field or this would create additional obstacles to the growth of the field. Already there have been governance implications for biosafety and biosecurity, as it "remains an extension of recombinant DNA technology and the scientific community commits to developing voluntary codes of conduct" [69]. The EC and member states should support education and training programs of SB, while the societal and scientific community should be involved in the continuing debate to balance the self-governance and regulation. It was also suggested that the EC should build a robust governance framework and raise the governance issues internationally, particularly in the areas of research funding, ethics and human rights, and biosecurity, as well as trade and IPR [68]. It is believed that the right governance tools will help the responsible use of SB to promote scientific advances that would benefit the whole society and the environment.

A report from a workshop organized by ERASynBio on public dialogue and governance suggested that governance of SB should be based particularly on three principles: participation, transparency, and accountability (see http://www.erasynbio.eu). These principles should then be implemented at all levels of the ERA-net – from strategies to individual projects. These principles should be reflected in the calls and in the evaluation processes. The EC expert group provided opinions on how to implement RRI regarding to governance aspect (as listed in Box 18.1):

To enable continuing RRI, the policy makers have called for collaboration with all stakeholders. This includes calling from funders on collaborative projects for researchers from the natural and social sciences. The convergence of both science aims to enhance both the scientific quality and the extent how social and ethical considerations are integrated [96]. The expert group of EC suggested that the societal stakeholders should be not only get involved in the projects but also get involved in the funding evaluation processes [65].

Box 18.1 Options to implement RRI [65]

- Applicants for EU research funds have to submit a statement on the ethical aspects of their research. This could be emphasized and applied more broadly. Additional guidance could be offered to applicants on the completion of this section.
- Asking for a statement in each research proposal on how the research might contribute to addressing societal challenges (similar to the outline on the consideration of the Gender dimension)
- The potential contributions to societal needs and the consideration of ethical aspects could become part of the selection criteria for research projects. So far, proposals are assessed against (i) scientific excellence, (ii) potential impacts (broadly defined), and (iii) management of the project. RRI aspects could be considered as a fourth aspect or a specification of the potential impacts.

18.5 Conclusion

SB is a nascent and innovative field of research with the potential to contribute to the whole society by addressing some of the challenges we are facing today, ranging from sustainable energy to green economy to environmental remediation. The industrial potential is believed to be huge, and many scientists, politicians, and industrialists see SB as the key to the knowledge-based bioeconomy. Right now the public knows little about SB, and the public awareness of the field is growing at a slow speed as indicated by the studies on the public attitude toward SB in Europe and in the United States. With the European conflict on the use of GM crops still presents in many people's minds, some fear that SB could run into similar problems, seeing SB as a mere GE 2.0, thus halting the development process. Other observers underline the interdisciplinary character of SB, pointing out that it might be a real converging technology where nanotechnology, IT, and biotechnology all come together. What is true for the technological convergence could also be true for the public debate about the technology, of which we have elucidated in the frames used in the public debate and the comparators of SB. Since SB is as much a GE as it is a form of nanotechnology and a form of IT, the respective public image will influence the developing public perception of SB. While GE contributes with the "technology as conflict" comparator, nanotechnology is represented in the notion of "technology as progress," and IT is "technology as gadget." As a consequence, SB might not be a GE 2.0 but something different, something new and unique.

Findings from the studies on the public perception of SB as well as the analysis on the frames and comparators of SB call for an innovation approach to address these issues. SB, at least in Europe, should be developed along with an attempt to promote a contemporary approach to technological development – RRI. This approach takes up issues of stakeholder participation, science education, gender equality, open access, ethics, and governance and can be seen as a comprehensive approach to deal with novel technologies in an environment that does not

automatically praise every scientific development per se and a priori as great, but asks how the technology could not only support the (economy) but actually benefit people and the environment.

Learning from the lesson of the past, applying the RRI approach in SB is critical to facilitate the social benefits of an emerging technology, to avoid raising social resistance to a technological advance that does not benefit people, and to support trust in research and innovation. More activities to get the public involved should be encouraged, while novel models should be built for open dialogues on SB. One of our ongoing projects – SYNENERGENE – may provide such models by setting up six platforms to tackle issues and challenges on SB of future of the field, public science and participation, art, culture and society, research and policy, international dimension, and online communication. We also expect to see more activities on turning SB into an RRI in the coming years.

Acknowledgments

Author acknowledges the financial support of the EC-FP7 Project ST-FLOW (EC Grant No. 289326) and SYNENERGENE (EC Grant No. 321488) as well as Project SYNMOD (FWF Grant No. I490-B12).

References

1 PCSBI (2010) New Directions: the Ethics of Synthetic Biology and Emerging Technologies, http://bioethics.gov/synthetic-biology-report.
2 Purnick, P.E. and Weiss, R. (2009) The second wave of synthetic biology: from modules to systems. *Nat. Rev. Mol. Cell Biol.*, **10** (6), 410–422.
3 IRGC (2008) Synthetic Biology: Risk and Opportunities of an Emerging Field, http://www.synbiosafe.eu/uploads///pdf/IRGC_ConceptNote_SyntheticBiology_Final_30April.pdf (accessed 30 September 2017).
4 Schwille, P. (2011) Bottom-up synthetic biology: engineering in a tinkerer's world. *Science*, **333** (6047), 1252–1254.
5 Berg, P., Baltimore, D., Brenner, S., Roblin, R.O., and Singer, M.F. (1975) Summary statement of the asilomar conference on recombinant DNA molecules. *Proc. Natl. Acad. Sci. U.S.A.*, **72**, 1981–1984.
6 Keasling, J.D. (2008) Synthetic biology for synthetic chemistry. *ACS Chem. Biol.*, **3** (1), 64–76.
7 Juhas, M., Eberl, L., and Glass, J.I. (2011) Essence of life: essential genes of minimal genomes. *Trends Cell Biol.*, **21** (10), 562–568.
8 Luisi, P.L. (2007) Chemical aspects of synthetic biology. *Chem. Biodivers.*, **4** (4), 603–621.
9 Murtas, G. (2007) Question 7: construction of a semi-synthetic minimal cell: a model for early living cells. *Origins Life Evol. Biosphere*, **37** (4–5), 419–422.
10 Bedau, M.A., Parke, E.C., Tangen, U., and Hantsche-Tangen, B. (2009) Social and ethical checkpoints for bottom-up synthetic biology, or protocells. *Syst. Synth. Biol.*, **3** (1–4), 65–75.

11 Giordano-Coltart, J. and Calkins, C.W. (2012) Best practices in patent license negotiations. *Nat. Biotechnol.*, **25**, 1–3.

12 Ackermann, D., Schmidt, T.L., Hannam, J.S., Purohit, C.S., Heckel, A., and Famulok, M. (2010) A double-stranded DNA rotaxane. *Nat. Nanotechnol.*, **5** (6), 436–442.

13 Loakes, D. and Holliger, P. (2009) Darwinian chemistry: towards the synthesis of a simple cell. *Mol. BioSyst.*, **5** (7), 686.

14 Marliere, P., Patrouix, J., Doring, V., Herdewijn, P., Tricot, S., Cruveiller, S., Bouzon, M., and Mutzel, R. (2011) Chemical evolution of a bacterium's genome. *Angew. Chem. Int. Ed.*, **50** (31), 7109–7114.

15 Pinheiro, V.B. and Holliger, P. (2012) The XNA world: progress towards replication and evolution of synthetic genetic polymers. *Curr. Opin. Chem. Biol.*, **16** (3–4), 245–252.

16 Pinheiro, V.B., Loakes, D., and Holliger, P. (2013) Synthetic polymers and their potential as genetic materials. *BioEssays*, **35** (2), 113–122.

17 Pinheiro, V.B., Taylor, A.I., Cozens, C., Abramov, M., Renders, M., Zhang, S., Chaput, J.C., Wengel, J., Peak-Chew, S.Y., McLaughlin, S.H., Herdewijn, P., and Holliger, P. (2012) Synthetic genetic polymers capable of heredity and evolution. *Science*, **336** (6079), 341–344.

18 Hart Research Associates (2013) *Awareness & Impressions of Synthetic Biology*, Synthetic Biology Project, Washington, DC.

19 Connor, M.R. and Atsumi, S. (2010) Synthetic biology guides biofuel production. *J. Biomed. Biotechnol.*, **2010**, 1–9.

20 Dellomonaco, C., Fava, F., and Gonzalez, R. (2010) The path to next generation biofuels: successes and challenges in the era of synthetic biology. *Microb. Cell Fact.*, **9** (1), 3.

21 Sommer, M.O., Church, G.M., and Dantas, G. (2010) A functional metagenomic approach for expanding the synthetic biology toolbox for biomass conversion. *Mol. Syst. Biol.*, **6**, 360.

22 Gibson, D.G., Glass, J.I., Lartigue, C., Noskov, V.N., Chuang, R.Y., Algire, M.A., Benders, G.A., Montague, M.G., Ma, L., Moodie, M.M., Merryman, C., Vashee, S., Krishnakumar, R., Assad-Garcia, N., Andrews-Pfannkoch, C., Denisova, E.A., Young, L., Qi, Z.Q., Segall-Shapiro, T.H., Calvey, C.H., Parmar, P.P., Hutchison, C.A. III, Smith, H.O., and Venter, J.C. (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. *Science*, **329** (5987), 52–56.

23 Navid, E.L. and Einsiedel, E.F. (2012) Synthetic biology in the Science Café: what have we learned about public engagement? *J. Sci. Commun.*, **11** (4), 1–9.

24 Pauwels, E. (2009) Review of quantitative and qualitative studies on U.S. public perceptions of synthetic biology. *Syst. Synth. Biol.*, **3** (1–4), 37–46.

25 Pauwels, E. and Ifrim, I. (2008) *Trends in American and European Press Coverage of Synthetic Biology*, Synthetic Biology Project, Washington, DC.

26 Gillespie, I.M. and Philp, J.C. (2013) Bioremediation, an environmental remediation technology for the bioeconomy. *Trends Biotechnol.*, **31** (6), 329–332.

27 European Commission (2010) Europeans and Biotechnology in 2010 Winds of Change? http://ec.europa.eu/research/science-society/document_library/pdf_06/europeans-biotechnology-in-2010_en.pdf (accessed 30 September 2017).

28 TNS Opinion & Social (2010) Eurobarometer 73.1 Biotechnology, http://ec.europa.eu/public_opinion/archives/ebs/ebs_341_en.pdf (accessed 30 September 2017).

29 Gaskell, G., Allansdottir, A., Allum, N., Castro, P., Esmer, Y., Fischler, C., Jackson, J., Kronberger, N., Hampel, J., Mejlgaard, N., Quintanilha, A., Rammer, A., Revuelta, G., Stares, S., Torgersen, H., and Wager, W. (2011) The 2010 Eurobarometer on the life sciences. *Nat. Biotechnol.*, **29** (2), 113–114.

30 SCENIHR, SCHER and SCCS (2015) Opinion on Synthetic Biology III – Risks to the Environment and Biodiversity Related to Synthetic Biology and Research Priorities in the Field of Synthetic Biology, http://ec.europa.eu/health/scientific_committees/emerging/docs/scenihr_o_050.pdf

31 SCHER, SCENIHR and SCCS (2014) Opinion on Synthetic Biology I – Definition, http://ec.europa.eu/health/scientific_committees/emerging/docs/scenihr_o_044.pdf

32 SCHER, SCENIHR and SCCS (2015) Opinion on Synthetic Biology II – Risk Assessment Methodologies and Safety Aspects, http://ec.europa.eu/health/scientific_committees/consultations/public_consultations/scenihr_consultation_26_en.htm

33 Kwok, R. (2012) DNA's new alphabet. *Nature*, **491**, 516–518.

34 Kronberger, N., Holtz, P., and Wagner, W. (2011) Consequences of media information uptake and deliberation: focus groups' symbolic coping with synthetic biology. *Public Understanding Sci.*, **21** (2), 174–187.

35 Cserer, A. and Seiringer, A. (2009) Pictures of synthetic biology: a reflective discussion of the representation of synthetic biology (SB) in the german-language media and by SB experts. *Syst. Synth. Biol.*, **3** (1–4), 27–35.

36 Hauser, J. and Schmidt, M. (2011) Synthethic, http://www.biofaction.com/pdf/gallery_guide.pdf

37 Kerbe, W. and Schmidt, M. (2015) Splicing boundaries: the experiences of bioart exhibition visitors. *Leonardo*, **48** (2), 128–136.

38 DFG, Acatech and Lepoldina (2009) Synthetic Biology: Positions, http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2009/stellungnahme_synthetische_biologie.pdf

39 Pei, L., Gaisser, S., and Schmidt, M. (2012b) Synthetic biology in the view of European public funding organisations. *Public Understanding Sci.*, **21** (2), 149–162.

40 Mampuys, R. and Brom, F.W.A. (2010) The quiet before the storm: anticipating developments in synthetic biology. *Poiesis Prax.*, **7** (3), 151–168.

41 Est, R.V., Vriend, H.D., and Walhout, B. (2007) Constructing Life: the World of Synthetic Biology, http://www.rathenau.nl/en/publications/publication/constructing-life-1.html

42 RAE (2009) Synthetic Biology: Public Dialogue on Synthetic Biology, http://www.raeng.org.uk/news/publications/list/reports/Syn_bio_dialogue_report.pdf

43 Bhattachary, D., Calitz, J.P., and Hunter, A. (2010) Synthetic Biology Dialogue, http://www.bbsrc.ac.uk/web/FILES/Reviews/1006-synthetic-biology-dialogue.pdf

44 BBSRC, EPSRC and Sciencewise-ERC (2011) Synthetic Biology Dialogue – Summary Report, http://www.bbsrc.ac.uk/web/FILES/Reviews/synbio_summary-report.pdf

45 Stemerding, D., Vriend, H.D., Walhout, B., and Est, R.V. (2009) Synthetic biology and the role of civil society rrganizations, in *Synthetic Biology* (ed. M. Schmidt), Springer Science+Business Media B.V.

46 Bereswill, S., Munoz, M., Fischer, A., Plickert, R., Haag, L.M., Otto, B., Kuhl, A.A., Loddenkemper, C., Gobel, U.B., and Heimesaat, M.M. (2010) Anti-inflammatory effects of resveratrol, curcumin and simvastatin in acute small intestinal inflammation. *PLoS One*, **5** (12), e15099.

47 ETC (2007) Extreme Genetic Engineering an Introduction to Synthetic Biology, http://www.etcgroup.org/content/extreme-genetic-engineering-introduction-synthetic-biology

48 ETC, Friends for the Earth and CTA (2012) The Principles for the Oversight of Synthetic Biology, http://www.synbioproject.org/process/assets/files/6620/_draft/principles_for_the_oversight_of_synthetic_biology.pdf

49 ETC and Heinrich Böll Foundation (2012) Biomassters Battle to Control the Green Economy, http://www.etcgroup.org/sites/www.etcgroup.org/files/greco_A4_eng_v16.pdf

50 ETC (2006) Global Coalition Sounds the Alarm on Synthetic Biology, http://www.etcgroup.org/content/global-coalition-sounds-alarm-synthetic-biology

51 ETC (2010) The New Biomassters, http://www.etcgroup.org/content/new-biomassters

52 Bogner, A. (2010) Let's disagree! Talking ethics in technology controversies. *Sci. Technol. Innov. Stud.*, **6** (2), 183–201.

53 Simmel, G. (1958) *Der Streit. Soziologie. Untersuchungen über die Formen der Vergesellschaftung*, Duncker & Humblot, Berlin.

54 Moya, A., Krasnogor, N., Pereto, J., and Latorre, A. (2009) Goethe's dream. Challenges and opportunities for synthetic biology. *EMBO Rep.*, **10** (Suppl. 1), S28–S32.

55 Ulsemer, P., Toutounian, K., Schmidt, J., Leuschner, J., Karsten, U., and Goletz, S. (2012) Safety assessment of the commensal strain Bacteroides xylanisolvens DSM 23964. *Regul. Toxicol. Pharmacol.*, **62** (2), 336–346.

56 Torgersen, H. and Schmidt, M. (2013) Frames and comparators: how might a debate on synthetic biology evolve? *Futures*, **48** (100), 44–54.

57 Lindgren, S. (1999) Biosafety aspects of genetically modified lactic acid bacteria in EU legislation. *Int. Dairy J.*, **9**, 37–41.

58 Torgersen, H. and Schmidt, M. (2012) Perspektiven der Kommunikation für die Synthetische Biologie, in *Biotechnologie-Kommunikation. Kontroversen, Analysen, Aktivitäten*, Acatech Diskussion (eds M.-D. Weitze, A. Pühler, W.M. Heckl, *et al.*), Springer-Verlag, Heidelberg, p. 113.

59 IRGC (2010) Guidelines for the Appropriate Risk Governance of Synthetic Biology, www.irgc.org/IMG/pdf/irgc_SB_final_07jan_web.pdf

60 Grobe, A., Eberhard, C., and Hutterli, M. (2005) *Nanotechnologie im Spiegel der Medien: Medienanalyse zur Berichterstattung uber Chancen und Risiken der Nanotechnologie*, Stiftung Risiko-Dialog, St. Gallen.

61 Sybesma, W., Hugenholtz, J., De Vos, W.M., and Smid, E.J. (2006) Safe use of genetically modified lactic acid bacteria in food. Bridging the gap between consumers, green groups, and industry. *Electron. J. Biotechnol.*, **9** (4).

62 von Schomberg, R. (2012) Prospects for technology assessment in a framework of responsible research and innovation, in *Technikfolgen abschätzen lehren* (eds M. Dusseldorp and R. Beecroft), Fachmedien Wiesbaden, Springer, pp. 39–61.

63 Schmidt, M., Meyer, A., and Cserer, A. (2015) The film festival bio:fiction: sensing possibilities how a debate about synthetic biology might evolve. *Public Understanding Sci.*, **24** (5), 619–635.

64 UK Synthetic Biology Roadmap Coordination Group (2012) A Synthetic Biology Roadmap for the UK, www.innovateuk.org/_assets/tsb_syntheticbiologyroadmap.pdf

65 European Commission (2013) Options for Strengthening Responsible Research and Innovation, http://ec.europa.eu/research/science-society/document_library/pdf_06/options-for-strengthening_en.pdf

66 European Commission (2012a) Mobilisation and Mutual Learning (MML) Action Plans on Societal Challenges, http://ec.europa.eu/research/science-society/document_library/pdf_06/mobilisation-mutual-learning-work-programme-2012_en.pdf

67 SYNENERGENE (2013) http://www.synenergene.eu/ (accessed 30 September 2017).

68 EGE (2009) Ethics of Synthetic Biology, http://ec.europa.eu/bepa/european-group-ethics/docs/opinion25_en.pdf

69 EASAC (2010) Realising European Potential in Synthetic Biology: Scientific Opportunities and Good Governance, http://www.cbd.int/doc/emerging-issues/emergingissues-2013-10-EASAC-SyntheticBiology-en.pdf

70 Technology Strategy Board (2012) Responsible Innovation Framework for Commercialisation of Research Findings.

71 Sutcliffe, H. (2013) A Report on Responsible Research & Innovation, http://ec.europa.eu/research/science-society/document_library/pdf_06/rri-report-hilary-sutcliffe_en.pdf

72 Nerlich, B. and McLeod, C. (2016) The dilemma of raising awareness "responsibly": The need to discuss controversial research with the public raises a conundrum for scientists: when is the right time to start public debates? *EMBO Rep.*, **17** (4), 481–485.

73 European Commission (2012b) Structural Change in Research Institutions: Enhancing Excellence, Gender Equality and Efficiency in Research and Innovation, http://ec.europa.eu/research/science-society/document_library/pdf_06/structural-changes-final-report_en.pdf (accessed 30 September 2017).

74 Sybhel (2012) Synthetic Biology & Human Health: the Ethical and Legal Issues, http://sybhel.org/wp-content/uploads/2012/11/SYBHEL-Final-Report.pdf

75 Bruce, D. (2010) Playing Democs Games to Explore Synthetic Biology, http://www.edinethics.co.uk/synbio/synbio%20democs%20report.pdf

76 Council of the European Union (2012) Declaration of the European Ministers Responsible for the Integrated Maritime Policy and the European Commission, on a Marine and Maritime Agenda for Growth and Jobs the "Limassol Declaration", http://ec.europa.eu/maritimeaffairs/policy/documents/limassol_en.pdf

77 iGEM (2013) The iGEM, http://igem.org/Main_Page

78 Bennett, G., Gilman, N., Stavrianakis, A., and Rabinow, P. (2009) From synthetic biology to biohacking: are we prepared? *Nat. Biotechnol.*, **27** (12), 1109–1111.

79 Delgado, A. (2013) DIYbio: making things and making futures. *Futures*, **48**, 65–73.

80 DIYBio. (2013) An Institution for the Do-It-Yourself Biologist, http://diybio.org

81 Penders, B. (2011) DIY biology. *Nature*, **472**, 167.

82 Tocchetti, S. (2012) DIYbiologists as 'makers' of personal biologies: how MAKE magazine and maker faires contribute in constituting biology as a personal technology. *J. Peer Prod.*, (2), 1–8.

83 Guan, Z., Schmidt, M., Pei, L., Wei, W., and Ma, K. (2013) Biosafety considerations of synthetic biology in the international genetically engineered machine (iGEM) competition. *Bioscience*, **63** (1), 25–34.

84 Jefferson, C. (2013) Governing Amateur Biology: Extending Responsible Research and Innovation in Synthetic Biology to New Actors. Building a Sustainable Capacity in Dual-Use Bioethics, Wellcome Trust Project.

85 NSABB (2011) *Strategies to Educate Amateur Biologists and Scientists in Non-life Science Disciplines About Dual Use Research in the Life Sciences*, NSABB, Washington, DC.

86 Pei, L., Bar-Yam, S., Byers-Corbin, J., Casagrande, R., Eichler, F., Lin, A., Österreicher, M., Regardh, P.C., Turlington, R.D., Oye, K.A., Torgersen, H., Schmidt, M., Guan, Z.-J., and Wei, W. (2012a) Regulatory frameworks for synthetic biology, in *Synthetic Biology Industrial and Environmental Applications* (ed. M. Schmidt), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, p. 240.

87 Seyfried, G., Pei, L., and Schmidt, M. (2014) European do-it-yourself (DIY) biology: beyond the hope, hype and horror. *BioEssays*, **36** (6), 548–551.

88 Wohlsen, M. (2008) *Do it Yourself DNA: Amateurs Trying Genetic Engineering at Home*, The Huffington Post, http://www.huffingtonpost.com/2008/12/25/do-it-yourself-dna-amateu_n_153489.html

89 Saukshmya, T. and Chugh, A. (2010) Intellectual property rights in synthetic biology: an anti-thesis to open access to research? *Syst. Synth. Biol.*, **4** (4), 241–245.

90 Rai, A. and Boyle, J. (2007) Synthetic biology: caught between property rights, the public domain, and the commons. *PLoS Biol.*, **5** (3), e58.

91 BBF BioBricks Foundation, (2013) http://biobricks.org/about-foundation/

92 Ham, T.S., Dmytriv, Z., Plahar, H., Chen, J., Hillson, N.J., and Keasling, J.D. (2012) Design, implementation and practice of JBEI-ICE: an open source biological part registry platform and tools. *Nucleic Acids Res.*, **40** (18), e141.

93 European Commission (2007) Synthetic Biology: a NEST Pathfinder Initiative, ftp://ftp.cordis.europa.eu/pub/nest/docs/5-nest-synthetic-080507.pdf (accessed 30 September 2017).

94 Synbiosafe Safety and Ethical Aspects of Synthetic Biology, http://www.synbiosafe.eu (accessed 30 September 2017).

95 Gutmann, A. (2011) The ethics of synthetic biology: guiding principles for emerging technologies. *Hastings Cent. Rep.*, **41** (4), 17–22.

96 Flipse, S.M., van der Sanden, M.C., and Osseweijer, P. (2013) Setting up spaces for collaboration in industry between researchers from the natural and social sciences. *Sci. Eng. Ethics*, **20**, 7–22.

Index

Synthetic Biology: Parts, Devices and Applications, First Edition. Edited by Christina Smolke.

d

e

f

g