Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

FROM GENOME TO THERAPY: INTEGRATING NEW TECHNOLOGIES WITH DRUG DEVELOPMENT

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

The Novartis Foundation is an international scientific and educational charity (UK Registered Charity No. 313574). Known until September 1997 as the Ciba Foundation, it was established in 1947 by the CIBA company of Basle, which merged with Sandoz in 1996, to form Novartis. The Foundation operates independently in London under English trust law. It was formally opened on 22 June 1949.

The Foundation promotes the study and general knowledge of science and in particular encourages international co-operation in scientific research. To this end, it organizes internationally acclaimed meetings (typically eight symposia and allied open meetings and 15–20 discussion meetings each year) and publishes eight books per year featuring the presented papers and discussions from the symposia. Although primarily an operational rather than a grant-making foundation, it awards bursaries to young scientists to attend the symposia and afterwards work with one of the other participants.

The Foundation's headquarters at 41 Portland Place, London W1N 4BN, provide library facilities, open to graduates in science and allied disciplines. Media relations are fostered by regular press conferences and by articles prepared by the Foundation's Science Writer in Residence. The Foundation offers accommodation and meeting facilities to visiting scientists and their societies.

Information on all Foundation activities can be found at
http://www.novartisfound.org.uk

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

Novartis Foundation Symposium 229

FROM GENOME TO THERAPY: INTEGRATING NEW TECHNOLOGIES WITH DRUG DEVELOPMENT

2000

JOHN WILEY & SONS, LTD

Chichester · New York · Weinheim · Brisbane · Singapore · Toronto

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

Published in 2000 by John Wiley & Sons Ltd,
Baffins Lane, Chichester,
West Sussex PO19 1UD, England

National 01243 779777
International (+44) 1243 779777
e-mail (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on http://www.wiley.co.uk
or http://www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons, Inc., 605 Third Avenue,
New York, NY 10158-0012, USA

WILEY-VCH Verlag GmbH, Pappelallee 3,
D-69469 Weinheim, Germany

Jacaranda Wiley Ltd, 33 Park Road, Milton,
Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01,
Jin Xing Distripark, Singapore 129809

John Wiley & Sons (Canada) Ltd, 22 Worcester Road,
Rexdale, Ontario M9W 1L1, Canada

Novartis Foundation Symposium 229
viii+165 pages, 21 figures, 7 tables

Library of Congress Cataloging-in-Publication Data

From genome to therapy : integrating new technologies with drug development / [editors: Gregory R. Bock, Dalia Cohen, and Jamie A. Goode].
p. cm. – (Novartis Foundation symposium ; 229)
"Symposium on From Genome to Therapy: Integrating New Technologies with Drug Development, held at the Hotel Europe, Basel, Switzerland, 22–24 June 1999"–Contents p.
Includes bibliographical references and index.
ISBN 0-471-62744-5 (alk. paper)
1. Pharmacogenomics–Congresses. 2. Pharmacogenetics–Congresses. I. Bock, Gregory. II. Cohen, Dalia, Ph.D. III. Goode, Jamie. IV. Novartis Foundation. V. Symposium on From Genome to Therapy: Integrating New Technologies with Drug Development (1999 : Basel, Switzerland) VI. Series.
RM 301.3.G45 F76 2000
615′.19–dc21 00-043348

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0 471 62744 5

Typeset in 12 on 14 pt Garamond by Dobbie Typesetting Limited, Tavistock, Devon.
Printed and bound in Great Britain by Biddles Ltd, Guildford and King's Lynn.
This book is printed on acid-free paper responsibly manufactured from sustainable forestry, in which at least two trees are planted for each one used for paper production.

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

Contents

Symposium on From genome to therapy: integrating new technologies with drug development, held at the Hotel Europe, Basel, Switzerland, 22–24 June 1999

Editors: Gregory R. Bock, Dalia Cohen (Organizers) and Jamie A. Goode

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

Participants

Allan Bradley Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

Steve D. M. Brown MRC Mammalian Genetics Unit, Harwell, Oxfordshire OX11 0RD, UK

Dalia Cohen Head of Functional Genomics, Novartis Pharmaceutical Corporation, Room LSB 1237, 556 Morris Avenue, Summit, NJ 07901, USA

J.William Efcavitch Applied Biosystems, 850 Lincoln Centre Drive, Foster City, CA 94404-1128, USA

Claire M. Fraser The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

Peter N. Goodfellow SmithKline Beecham Pharmaceuticals, New Frontiers Science Park, Third Avenue, Harlow, Essex CM19 5AW, UK

Richard Goold Incyte Pharmaceuticals Inc., 3174 Porter Drive, Palo Alto, CA 94304, USA

Paul L. Herrling Head of Research, Novartis Pharma AG, CH-4002 Basel, Switzerland

Denis Hochstrasser Laboratoire Central de Chimie Clinique, Hôpital Cantonal Universitaire, 24 rue Micheli-du-Crest, CH-1211 Geneva, Switzerland

Stephen L. Hoffman Naval Medical Research Center, 503 Robert Grant Avenue, Silver Spring, MD 20910, USA

Jenny Kopczynski Genetics, Exelixis Pharmaceuticals, 260 Littlefield Avenue, South San Francisco, CA 94080, USA

Robert J. Lipshutz Affymetrix Inc., 3380 Central Expressway, Santa Clara, CA 95051, USA

David Magnus University of Pennsylvania Center for Bioethics, 3401 Market Street #320, Philadelphia, PA 19104-3308, USA

Matthias Mann Protein Interaction Laboratory, University of Southern Denmark–Odense, Campusvej 55, DK-5230 Odense M, Denmark

Saira Mian (*Novartis Foundation bursar*) Life Sciences Division (Mail Stop 29-100), Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

Allen Roses Glaxo Wellcome Research and Development, 5 Moore Drive, Research Triangle Park, NC 27709, USA

Gerald M. Rubin University of California at Berkeley, Department of Molecular & Cell Biology, 142 Life Sciences Addition #3200, Berkeley, CA 94720-3200, USA

Larry M. Souza AMGEN, Inc., One Amgen Center, Thousand Oaks, CA 91320, USA

Joseph Straus Max-Planck-Institute for Foreign and International Patent, Copyright and Competition Law, Marstallplatz 1, D-80539 Munich, Germany

Jan van Oostrum Head of Protein Sciences, Functional Genomics Area, Novartis Pharma AG, CH-4002 Basel, Switzerland

J. Craig Venter (*Chairman*) Celera Genomics Corporation, 45 West Gude Drive, Rockville, MD 20850, USA

Elizabeth Winzeler[1] Department of Biochemistry, Stanford University School of Medicine, Stanford, CA 94305-5307, USA

[1]Current address: Genomics Institute of the Novartis Research Foundation, 3115 Merryfield Row, Suite 200, San Diego, CA 92121, USA

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

Chairman's introduction

J. Craig Venter

Celera Genomics Corporation, 45 West Gude Drive, Rockville, MD 20850, USA

It is exciting that the Novartis Foundation 50th Anniversary meeting is being held on genomics. Claire Fraser and I participated in the Novartis (then Ciba) Foundation symposium on Receptors, antibodies and disease in 1981—we actually spent our honeymoon there. Between attending the conference sessions we had to write a research grant in the Foundation's library.

Let me introduce the topic. A book I read recently described the first statistical survey on death rates and longevity. This dates from London in the 15th century and showed the percentages of people that were alive at different ages. By age 46 only about 10% of the initial population was still alive. This high death rate was largely due to infectious diseases, primarily the plague. Those of us here who are in our 50s would represent an absolute minority of the population. The fact that now 70% of the population in western nations is still alive in their 70s shows the impact of science and technology.

Back in 1990 we tried to organize a genomics meeting at one of the first genome conferences, and there was very little interest. At that time most people in the pharmaceutical industry thought that genomics had no impact on what they did and that it was part of some futuristic hope that might happen at some point down the road. In 1990 about 1000 human genes were known; Genbank was incredibly small. Genomics began to have an impact with the advent of expressed sequence tags (ESTs), in 1991. ESTs now represent over 70% of Genbank accessions. When two companies, Incyte and Human Genome Sciences (HGS), began the effective commercialization of ESTs, this acted as a wake up call to the pharmaceutical industry. Incyte in particular has done a good job of making this data available and useful for industry. This initiated a real

change in the use of genomics in the development of drugs. This was an early stage in the development of genomics.

The field took another leap, starting in the middle of the 1990s, when the convergence of mathematics and ESTs allowed us to complete the first genome—that of *Haemophilus influenzae*. There has been a great change in the number of genes available since that time. We are in the early part of an exponential growth phase in which genomes of all types are being deciphered. This is important in context of the original planning of the genome project (at least in the USA) where they decided that there were only five organisms that needed to have their genomes decoded to provide a basis for all life. This included *Escherichia coli*, *Saccharomyces cerevisiae* (the fourth genome completed, and the first of a eukaryote), *Drosophila*, mouse and human. *Caenorhabditis elegans* was added early on. Despite this early narrow view of what would be informative, the current list of organisms whose genomes have been (or are being) sequenced is growing substantially. At The Institute for Genomic Research world wide web site (*http://www.tigr.org*), these are listed in order. The number of completed genomes that have so far been published in the scientific literature is close to two dozen. Looking down the list of genomes that are in progress (about 100), one can see many key pathogens. Some genomes are being done on a distributive model, where several institutions are participating. We will hear later from Steve Hoffman on the *Plasmodium* genome, where the 30 Mb genome has been broken up into different bits which are being sequenced at separate sites. Clearly, there is an ongoing tidal wave of data coming out of this work, which presents several challenges to the research community.

One of the facts that will prove to be the most important challenge to all of us is that roughly half of the genes in each species that have been sequenced are completely new, so far unknown genes, which seem to be species specific. The other half are highly conserved, and are seen in many species. During this meeting we will hear about techniques to meet the challenge of understanding the biology associated with the 50% of genes that we haven't seen before. With humans, the number of unknown genes is even higher.

Another issue I hope we will address during this symposium is the changing central dogma that one gene leads to one transcript which leads to one protein with a single function. There is now consensus that

the central dogma is more complicated. Multiple transcripts, different splice variants, RNAs coming out of introns and other regions for regulation, and multiple different protein forms with complex functions have made us adopt a more complex view. One of the best examples of this is with the so-called cystic fibrosis gene. In 1989, Francis Collins isolated the chloride ion channel that was linked to cystic fibrosis. Up until one year ago, if you had asked anyone what mutations in this gene would have led to, the universal answer would have been cystic fibrosis. From studies published recently in the *New England Journal of Medicine*, it is clear that mutations and spelling variations in this one gene can lead to a wide variety of medical outcomes (Cohn et al 1998, Sharer et al 1998). Changes can lead to chronic pancreatitis, asthma, male sterility or full-blown cystic fibrosis. More disturbing for most people, changes in this gene can lead to no apparent illness whatsoever. This notion of genetic determinism in an absolute sense is in need of serious re-thinking. However, we should not find the more complex notion surprising: we have one hundred trillion cells in our bodies and around 100 000 genes, all changing dynamically through development. So it is not inconceivable that one gene product can have cellular interactions with a wide variety of outcomes. This will be one of the challenges we must address as we move forward in genomics and genetics. This affects how we think about both diagnostics and treatment. If your job is to come up with a new drug to treat this disease and all the focus is on one protein, this is of crucial importance. If it is in diagnostics, countless pregnancies have been terminated because the fetus was tested and found to have changes in the chloride ion channel, which people were absolutely certain was going to lead to cystic fibrosis.

I hope that during this meeting we will hear of approaches and techniques that will help us understand the genome, and how the application of this insight can lead to new forms of therapy.

References

Cohn JA, Friedman KJ, Noone PG, Knowles MR, Silverman LM, Jowell PS 1998 Relation between mutations of the cystic fibrosis gene and idiopathic pancreatitis. N Engl J Med 339:653–658

Sharer N, Schwarz M, Malone G et al 1998 Mutations of the cystic fibrosis gene in patients with chronic pancreatitis. N Engl J Med 339:645–652

DISCUSSION

Cohen: I would like to ask you a question about the cystic fibrosis chloride ion channel mutations. The first mutations to be identified were large deletions at the 5′ end, and these were found to have dramatic effects, mostly on processing at the cell surface. Subsequently, other minor mutations have been found that are also clustered around the 5′ end. How do these latter mutations relate to the former, in terms of being responsible for different disease states?

Venter: I don't know the answer to your question. It is an example of the great complexity we're dealing with. Some people are trying to classify these mutations to determine whether different clusters of mutations are associated with different clinical states, but it's not yet clear if this is going to be the case. It is a feature of developmental biology that decisions are made constantly at different stages, and so minor aberrations in protein concentrations could have large impacts on developmental fate. This is a disturbingly complicated set-up, in terms of our goals of trying to intervene and correct developmental mistakes. It would be nice if there were clear-cut rules, such as 'changes in this gene always cause this particular disease state', but this is not the case. The assumptions so far have been that changes in the chloride ion channel always cause cystic fibrosis, but the new information coming out indicates that this is not the case, and people are going to have to think about these problems in a much broader sense. We are going to have to measure polymorphic variation in much broader population groups, rather than adopt a one-gene, one-disease approach.

Lipshutz: I agree. The cystic fibrosis investigations began by people who were looking at the cystic fibrosis genotype in individuals, rather than doing large-scale systematic screening of populations. Another area that remains largely unknown at present is how genes interact with each other, especially in terms of modifier genes and how these affect the manifestation of disease.

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

Electrophoresis-based fluorescent dideoxy-terminator sequencing

J. William Efcavitch

Applied Biosystems, 850 Lincoln Centre Drive, Foster City, CA 94070, USA

The introduction of real-time fluorescent dideoxy-terminator sequencing has enabled the bulk of the genome sequencing that has been performed over the past 10 years. Virtually all of these data have been acquired using an instrument system based on a batch process and a slab gel separation format (Connell et al 1987, Hood et al 1987). The demands for the acceleration of the finish to the sequencing of the human genome (Venter et al 1998), coupled with the increased use of genomics in the pharmaceutical discovery process has led to the recent development and introduction of production scale DNA analysers based on capillary electrophoresis. I will describe the current state of the art in fully automated DNA sequencing technology and additional technical advances which will continue to reduce the cost and increase the throughput of automated DNA sequencing.

Current instrumentation

In the fall of 1998, PE Biosystems introduced the Model 3700 DNA Analyzer, which uses 96 capillaries and sheath flow fluorescence detection to replace the slab gel and scanning fluorescence detector of previous instruments. Although this instrument system is mostly used in production-scale genome sequencing, it can also be used to perform high-throughput genotyping with *S*hort *T*andem *R*epeat markers. The electrophoretic separation is carried out in *c.* 50 μm diameter quartz capillaries, which are grouped in arrays of 96. These internally uncoated capillaries are filled with a non-cross-linked *N*,*N*-dimethylacrylamide polymer formulation that has been optimized for molecular weight, viscosity, denaturing properties and chemical stability. The hydrophobic

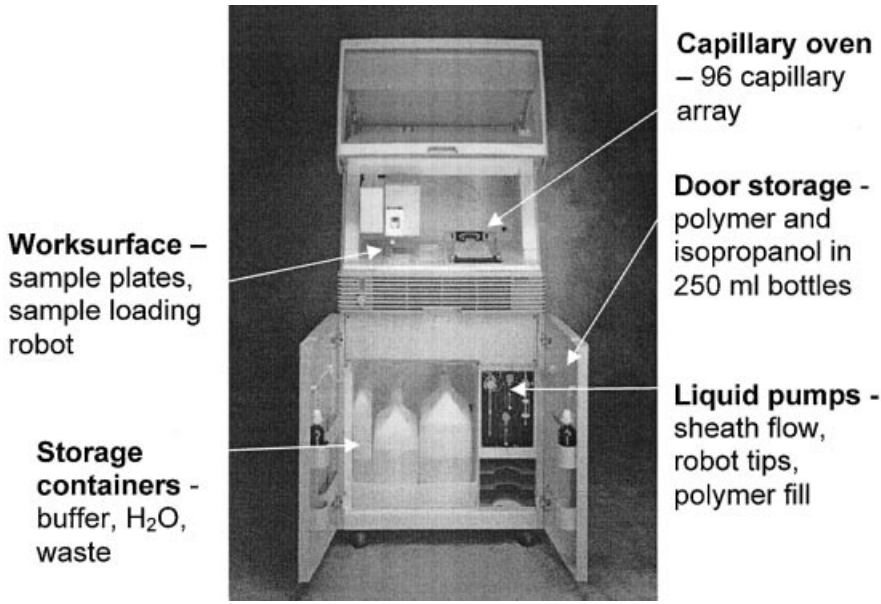


FIG. 1. PE Biosystems Model 3700 DNA Analyzer with lid and doors open.

polymer absorbs to the internal surface of the capillaries and suppresses the bulk liquid electroendosmotic flow and the interaction of the analytes with the capillary wall (Madabhushi et al 1996).

Twenty-four hour unattended operation of this capillary electrophoresis DNA analyser is achieved through the use of an integrated robotic pipetting system for sample introduction and a syringe pump system for replacement of the polymeric separation matrix between each electrophoretic analysis. A work surface which holds up to four 96-well or 384-well microwell plates, enables automatic access to the samples once the system has been properly configured via a computer workstation. Figure 1 shows an external view of the instrument highlighting the above systems.

Detection of the resolved dye terminator extension products occurs external to each of the capillaries after the fragments electrophoretically migrate out of the end of the capillaries and are transported by a low velocity fluid flow into the excitation zone of an Argon-ion laser. This detection process, called sheath flow detection, was utilized because of

TABLE 1 3700 DNA Analyzer performance

	Long sequencing	*Fast sequencing*	*Fragment analysis*
Length of read nt @ 98.5% accuracy	550–700+	350–450	500
Run time (h)	3.9	2.3–2.8	2.3
24 h throughput # lanes	768	864	864 or 12 960 genotypes

nt, nucleotides.

the sensitivity achievable in a 96-capillary format (Swerdlow et al 1990, Takahashi et al 1994).

Levels of performance for the current system are shown in Table 1. Performance changes such as electrophoresis speed or length of read are in general not limited by the hardware but are a function of the separation polymer formulation and running conditions. Reformulation of the separation polymer is currently underway and preliminary results indicate that run times of ~100 minutes should be achievable as shown in Fig. 2.

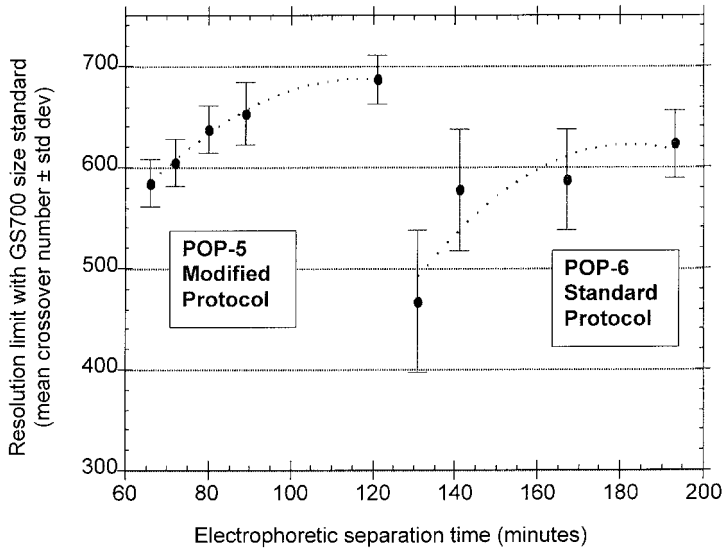


FIG. 2. 3700 DNA Analyzer performance increase as a function of polymer formulation. Electrophoretic resolution as a function of separation time for two different polymer formulations contained in the capillaries. Increase in both resolution limit and run speed can be influenced independent of the hardware configuration.

New developments

The cost, throughput and utility of electrophoresis-based DNA sequencing will continue to evolve as advances are made in the capillary electrophoresis separation process and some of the ancillary processes which are used to prepare the fluorescent-labelled dideoxy-terminator extension products prior to electrophoretic analysis.

ELFSE

The first area of innovation, which will greatly enhance the performance of electrophoresis-based DNA sequencing, is a separation principle called 'end-labelled free solution electrophoresis' (ELFSE), first described by Mayer et al (1994). Unlike classical gel electrophoresis, which is based upon resolution of the extension products by a sieving mechanism, ELFSE relies on a free solution separation of the fragments that vary by their charge to hydrodynamic friction ratio. Normally the electrophoretic mobility of nucleic acids is independent of charge and the hydrodynamic friction because their ratio is a constant for all fragment lengths. By attaching a drag-inducing label to the 5′ end of the sequencing primer, a free solution mechanism is enabled, since all of the fragments have the same friction coefficient but a different charge depending upon the number of nucleotides in each fragment (Fig. 3). Separations using this mechanism require that capillaries contain only a buffer and possibly a denaturant and that a wall coating suppresses the electroendosmotic flow. Since the separation is now gel-independent, the field strength can be increased dramatically to decrease the time-dependent diffusion. This system allows for either short sequence reads in tens to hundreds of seconds or possibly longer sequence reads than are currently achievable with conventional sieving systems.

Microfabricated microchannel electrophoresis

Hand in hand with the development of non-cross-linked, flowable capillary separation systems and the ELFSE separation system is the use of monolithic, microfabricated microchannel arrays to replace discrete capillary arrays (Manz et al 1992, Woolley & Mathies 1994). Since one of the resolution-limiting mechanisms in nucleic acid electrophoresis is Joule heating, which leads to band spreading, microchannels should

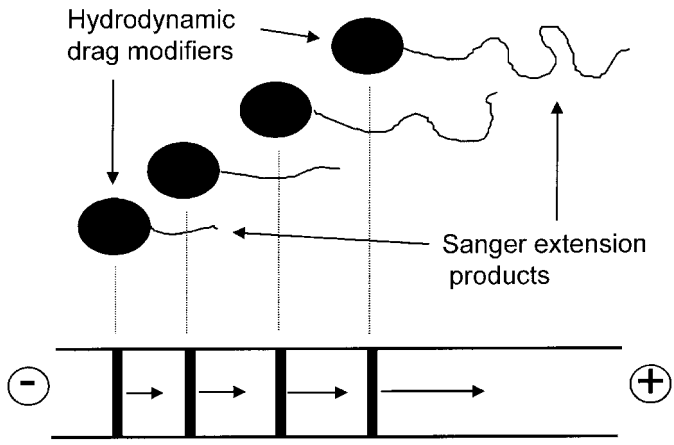


FIG. 3. Principle of end-labelled free solution electrophoresis (ELFSE). Sanger extension products are generated by conventional dye-dideoxynucleotide terminator reactions using a sequencing primer 5′-labelled with a hydrodynamic drag-inducing moiety such as streptavidin or a neutral peptide. Since each extension product is labelled with the same drag-inducing moiety, the mobility is differentially modulated by the number of negative charges associated with each differently sized extension product. Note that the longest fragment migrates the fastest since it bears the most charge, while the shortest fragment migrates the slowest since it bears the least charge.

allow performance gains by the fabrication of channel diameters smaller than practically achievable with discrete capillaries. In addition, controlled configurations of the injection zone for sample introduction should lead to additional performance gains that are not possible by simply dipping the end of a capillary into a sequencing reaction solution. This ability to use sub-microlitre sample sizes may require the integration of PCR amplification wells and reaction chambers at the head of each separation channel on the monolithic device. Such features could add to the cost of the devices but might ultimately reduce the labour and reagent cost of each sequencing analysis.

PCR-mediated template production

The process of dye-labelled dideoxynucleotide terminator sequencing still requires the preparation of template DNA from cell lysates. Although there exist a wide variety of solutions for the production of template

DNA from bacterial clones, one method which appears to be gaining acceptance is the use of PCR amplification of target insert DNA in plasmids from colony picks (Innis et al 1988, Gyllensten 1989). As the need for DNA sequencing moves from cloned-based *de novo* targets to whole genomic DNA isolates from routine or diagnostic samples, the robustness and simplicity of template production by PCR from cell lysates will lend itself to automation for high throughput analysis.

Multiplexed sequencing reactions and hybridization-based pullout

As the number of sequencing reactions grows exponentially, it becomes logical to seek methods for reducing some of the front-end labour associated with performing the dideoxy-terminator sequencing reactions themselves. In response to DNA sequencing moving from *de novo* sequencing of new genes to the comparative sequencing of mutations within known, whole genes, we have been developing a technology, called hybridization-based pullout (HBP), which will allow the simultaneous, one-tube cycle sequencing of many PCR amplicons (O'Neill et al 1998). Uniquely tailed sequencing primers will allow the sequential capture by hybridization of each individual sequencing ladder by separate solid supports. Captured fragments can be eluted and analysed by capillary electrophoresis DNA analysis. Satisfactory multiplex reactions, separation and analysis of up to 12 independent sequencing ladders has been demonstrated. HBP can be readily automated and, furthermore, could be incorporated into microfabricated microchannel electrophoresis devices.

Summary

Although electrophoretic-based DNA sequencing technology has been in place for more than 10 years, continued advances in the basic separation science, detection methodologies, automation and sample preparation promise to keep this technology in the forefront of genetic analysis. As the demands for sequence information moves from *de novo* whole genome analysis to more routine, comparative sequencing of known genes, we are confident that the technology will continue to evolve and will adapt to the demands of the scientific and commercial community.

Acknowledgements

I would like to thank the many individuals who have contributed to the development of the Model 3700 under the guidance of Michael Phillips and Kevin Hennessy. Additional thanks to Dave Hershey, Ben Johnson, Achim Karger and Roger O'Neill for specific contributions.

References

Connell C, Fung S, Heiner C et al 1987 Automated DNA sequence analysis. Biotechniques 5:342–348

Gyllensten UB 1989 PCR and DNA sequencing. Biotechniques 7:700–708

Hood LE, Hunkapiller MW, Smith LM 1987 Automated DNA sequencing and analysis of the human genome. Genomics 1:201–212

Innis MA, Myambo KB, Gelfand DH, Brow MA 1988 DNA sequencing with *Thermus aquaticus* DNA polymerase and direct sequencing of polymerase chain reaction-amplified DNA. Proc Natl Acad Sci USA 85:9436–9440

Madabhushi RS, Menchen SM, Efcavitch JW, Grossman PD 1996 Polymers for separation of biomolecules. US Patent 5,552,028

Manz A, Harrison DJ, Elisabeth MJ et al 1992 Planar chips technology for miniaturization and integration of separation techniques into monitoring systems. J Chromatogr 593:253–258

Mayer P, Slater GW, Drouin G 1994 Theory of DNA sequencing using free-solution electrophoresis of protein–DNA complexes. Anal Chem 66:1777–1778

O'Neill RA, Chen J-K, Chiesa C, Fry G 1998 Multiplex polynucleotide capture methods and compositions. WO 9814610A2

Swerdlow H, Wu S, Harke H, Dovichi NJ 1990 Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. J Chromatogr 516:61–67

Takahashi S, Murakami K, Anazawa T, Kambara H 1994 Multiple sheath-flow gel capillary-array electrophoresis for multicolor fluorescent DNA detection. Anal Chem 66:1021–1026

Venter JC, Adams MD, Sutton GG, Kerlavage AR, Smith HO, Hunkapiller M 1998 Shotgun sequencing of the human genome. Science 280:1540–1542

Woolley AT, Mathies RA 1994 Ultra-high-speed DNA fragment separations using microfabricated capillary array electrophoresis chips. Proc Natl Acad Sci USA 91:11348–11352

DISCUSSION

Venter: Sequencing capacity has doubled roughly every six months, whereas the costs have been progressively decreasing over this period. Many of the techniques that you have described, and particularly those pertaining to solution-based sequencing, have made this possible.

Rubin: Could you speculate what those sequencing costs will be in five years time? Will they continue to decrease at the same rate to, say, five cents or two cents per base?

Venter: It depends on how you calculate the cost. Currently costs are only 0.9 cents per base pair at the lowest-cost labs. If you include all the equipment costs, the overall sequencing costs are higher, but even so they will continue to fall. At the moment, the price of reagents is responsible for a large proportion of the costs. At Celera, we have been co-developing with PE Biosystems a multiplex sequencing analysis that uses the same number of reagents for two reactions. We are hoping to extend this to 10 reactions. This will decrease the cost per reaction by 10-fold. The ability to do electrophoresis in an aqueous solution has also substantially decreased the costs. The costs have been decreasing at such a rate that they have been causing unusual problems with institutions such as The Institute for Genomic Research (TIGR).

Fraser: Over the past year, the cost of sequencing the microbial genomes at TIGR has been reduced to approximately 17 cents per base pair. This dramatic decrease has resulted in grant monies going unspent. However, the good news is that much more sequence can be obtained today for the same cost as was required just one to two years ago.

Lipshutz: One of the questions that always comes to my mind is the trade-off between multiplexing and miniaturizing sample preparation reactions, including PCR. Where do you think some of these trade-offs are going to come from?

Efcavitch: I'm a little sceptical about the applications of miniaturization and associated integrated circuit technology to biological problems. This is long-term research, and we probably won't reap any benefits in the short term.

Lipshutz: What about simply trying to decrease the sample volumes?

Efcavitch: On paper, decreased sample volumes should lead to decreased costs; but these techniques will require complex devices and high development costs, and therefore they may not actually result in cost reductions.

Venter: We have been trying to push three issues to their limits without much success. These are evaporation, pipetting accuracy in the sub-nanolitre range and recovery of the sample. The challenge is to avoid developing expensive equipment that will offset any cost savings.

Mann: Bill Efcavitch, could you expand on your comments about free-flow electrophoresis and how do you see this technology progressing in the future?

Efcavitch: The principle is based on breaking the charge-to-mass relationship of nucleic acids. The sequencing primer is attached to a protein, or a large neutral species, which provides hydrodynamic drag. The charge modification is due to changes in nucleic acid length. At present, the limitation is the determination of the ideal mass and chemical properties of the drag modifier. The first studies were done with streptavidin and biotin, but we are now looking for larger synthetic peptides, because in order to have a read length of 500–600 nucleotides, it seems that a fairly large peptide is required, i.e. in the order of 100 kDa. We also know that the peptide must be neutral.

Mann: Would an optimum protein allow you to read out to 3 kb or more?

Efcavitch: I don't know. We would just be pleased if we could reach the existing read lengths, because this alone would decrease the running costs of the PE Biosystems 3700 sequencer. A read length of 3 kb is theoretically predicted, and is a target to aim for in the future. This is the most significant improvement in electrophoresis of nucleic acids in the last 30 years.

Venter: Is it dependent on a microchannel format, or does it also work on a macroscale?

Efcavitch: It works in a capillary, although to get to the ultimate performance, it will be necessary to control the initial zone width, which would probably require a microchannel format. All the microchannel work was originally done by people working on free-solution electrophoresis, in which diffusion and the initial zone width is critical for performance. Therefore, to use microchannels properly, we will have to resort to free-solution electrophoresis of nucleic acids.

Venter: What other sequencing techniques are emerging? Are people still working on enzymatic cleavage of single nucleotides, for example?

Efcavitch: I don't know of any groups that are continuing to work with this method.

Hochstrasser: What about mass spectrometry?

Efcavitch: Mass spectrometry is a strong player for high throughput short read sequencing and comparative sequencing, but probably not for long read *de novo* sequencing.

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

Genomic impact on pharmaceutical development

J. Craig Venter

Celera Genomics Corporation, 45 West Gude Drive, Rockville, MD 20850, USA

In the very near future, the development of important new pharmaceutical agents will undergo substantial changes. Pharmaceutical corporations willing to make the appropriate commitment can utilize the reference DNA sequence for the entire human genome, estimated to contain roughly 3.5 billion base pairs. They will have similar information on the entire genomes from a variety of model organisms essential to modern pharmaceutical development. Target discovery, lead compound identification, pharmacology, toxicology and clinical trials are likely to merge with the science of bioinformatics into a powerful system for developing new products. Genomic sequence databases, coupled with industrial-scale sequencing of full length cDNA, universal protein product libraries (proteomics) and the creation of powerful relational databases will transform the drug discovery process. There will be an ever-increasing opportunity for rational candidate drug design and the reduction of serious side effects. Research-based pharmaceutical companies will be able to use bioinformatics to analyse relevant gene classes and gene variations (polymorphisms), including the regulatory elements that govern gene expression. Comparative genomics and an analysis of synteny will allow dramatically more efficient prediction of gene structure and function. It will be possible to simulate the action of new molecules or therapeutic programs against diverse metabolic pathways, prior to pre-clinical testing. Thus, a paradigm of 'cyberpharmaceutical' testing will be available to the industry, speeding the selection of promising new agents, eliminating products that are likely to exhibit toxicity, and reducing the formidable costs and risks associated with the current paradigms of drug

development. The benefits to the pharmaceutical industry, and to the public served by this industry, will be incalculable, and are likely to emerge within the next decade.

DISCUSSION

Cohen: You mentioned in your talk that by sequencing the genome, you will be able to find a large number of single nucleotide polymorphisms (SNPs). How many individuals are you planning to sequence genes from in order to obtain enough information?

Venter: There are multiple questions embedded in your question. First, how do we generate a SNP database that represents the human population? At the moment, because of the scale of sequencing that would be required, sequencing individuals does not provide a complete answer to this question, although it is a starting point to create a database. It has been estimated that the five people and the 10 haplotypes that we intend to sequence, will provide a database of 80% of the abundant polymorphisms in the human population. But by definition they're abundant, and they will therefore have somewhat limited value. The answers will take a long time to work out. We need multiple approaches and techniques, which is where some of the high-throughput SNP technologies will be important. Ideally, we would like to have the sequence of everybody on the planet in a giant database, but we need an increase of few more orders of magnitude in the sequencing technologies in order to do this. The five individuals and 10 haplotypes is a starting point, but even this already exceeds our mathematical ability to organize the data: even the highest density Affymetrix chips can't measure 30 million polymorphisms as a data set. We therefore have to deal with limited data sets.

Cohen: This leads me to a more general question. When will you have enough genetic information in order to prescribe a medicine to a patient? I was asked this morning to guess whether clinical trials will be much shorter when we have more information about the polymorphic variation in individuals. This is a very important issue; we have to be very careful to give the right medicine to the right patient.

Venter: The early attempts are intended to provide a statistical paradigm, looking at a particular pattern in one individual versus

another pattern in a second individual. This tells us nothing about the individual effect of each of those variations. Ideally, for a knowledge base, at some point in the future, for every single polymorphic variation we measure there will be a phenotype or an outcome associated with that variation. This is not going to happen overnight: it is going to take decades or longer as we uncover the biology. If you had a complete database of all your variations versus the database of the complete sequence of everyone's genome, as each new discovery is made you just go and look that up and find its relevance to you. The broader screens early on allow data that we generate today, even though we don't totally understand their impact, to have a huge impact later on. It will progress from statistical paradigms to real knowledge-based research. I'm not sure that this is going to happen instantly in the clinical trial paradigm: it's going to take a long time to sort all this out.

Lipshutz: At a recent meeting on SNPs, Pui Kwok (personal communication) described how they are looking for SNPs in the overlapping bacterial artificial chromosomes (BACs) that are being sequenced in the publicly funded human genome project. They're looking at multiple individuals for the BAC clones. If they compare the BAC clone overlaps there are about 10 SNPs on each overlap.

Venter: It really depends on the BAC libraries. The initial absolute rules at NIH were that no more than 10% of the sequence could come from any one BAC library. However, because they didn't have BAC libraries of sufficient quality they just waived that rule, and now 40–50% of the sequences can come from one BAC library until they get some new ones made. I think that's why the two approaches, the whole genome shotgun method and the BAC methods, will actually be complementary. The genomes that are being done have come pretty much from a clonal set of information. Taking BACs and clones from different individuals with all the rearrangements we get in the human genome, it may be impossible to assemble a complete sequence from BACs alone.

Bradley: I have a question about the number of SNPs. Why do we need more SNPs than the number of genes? Obviously, you need different SNPs to define different ethnic groups, but why can't you just have one SNP per 3′ untranslated region?

Venter: It depends on what your goal is. I hope your goal in life isn't just to find genetic variation to identify ethnic groups. That would get us all in trouble pretty rapidly.

Bradley: So long as you can identify genetic variation in genes which affect phenoptypes, then this defines the number of SNPs that you need.

Lipshutz: You need to know the haplotypes, and there may be as many as 50 common haplotypes for any given region. If you think about how finely the genome has been divided up into different regions since our common ancestors, those regions may span over 30 or 40 kb for an average haplotype. Half a dozen may therefore be required for each 30–40 kb just to define the haplotypes, and that's not including the actual causative mutations. At the level of mapping and trying to do discovery, one per gene, if it is the right kind, may be sufficient. Most people would argue that more is better.

Venter: Especially if you want to understand protein function. We spent a lot of effort doing site-directed mutagenesis on seven-transmembrane receptors to try to understand variation in each amino acid and how it relates to function, and so there's a pretty good data set on these receptors in terms of variation at each site and what is really functional. If you have a SNP map with a lot of variation in those genes, you can go immediately from the sequence to trying to predict function in that individual. I think 30 million is going to be far too small a number on a genome-wide scale. Remember, only about 5% of those actually occur in genes, so the number may be right, they may just be in the wrong place. I would like to have a database that has all the variation in the human population in the regulatory regions of genes. Having tens and millions of them in other regions may be useful for some crude linkage disequilibrium studies, but not for relating genome sequence back to the protein function. So it all depends on whether the goal is mapping or trying to predict function.

Bradley: I think many mutations are ancient mutations, and the haplotypes are still quite large. Thus many polymorphisms will track with genes and will then track with disease. It depends whether the goal is to understand function or to have some association with disease so you can assign the right pharmaceutical treatment.

Venter: Hopefully, the goal is all the above. It seems that in a short while we should be able to overlay the chimpanzee genome on top of

the human genome. With mouse, we're going to overlay that on top of human and then we can start to understand some of the more ancient mutations. This will lead us to clues that may or may not help the pharmaceutical industry, but will certainly help us to understand what happened in evolution. The differences in people around the table in terms of the very rare alleles, could be the absolute key things for your uniqueness in the environment versus somebody else's. One example that is a fairly common allele is the one associated with aspirin sensitivity. That affects whether taking an aspirin a day helps you if you have a heart attack or stroke. It is just a single base pair variation which is possessed by one in three of the population. Currently, the way we practice medicine is that we tell everybody to take a baby aspirin each day, because we know it will affect one third of the population. This is the sort of instance where knowing the specific nucleotide variation is going to be very predictive in knowing pharmaceutical effects.

Mann: Can you give us an update on your policy regarding the availability of the Celera data?

Venter: Two groups represented here, Amgen and Novartis, get weekly updates by subscription. We are negotiating with a dozen major academic institutions in the USA for academic subscriptions. In terms of *Drosophila*, we indicated that once the *Drosophila* genome was completed, hopefully later in 1999, that we would be publishing the complete *Drosophila* genome sequence. Starting September 1, we plan to start adding to the Celera website availability of some of the *Drosophila* sequence increasing over time until the genome is totally completed, and that the basic sequence itself will be made available to academic institutions for no charge. But this will not be true for all the genomes that we do. There's a difference between free and accessible: in the US the *Wall Street Journal* is accessible to virtually everybody, but it's not free. Our goal is to have our data widely available and accessible, but not free.

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

From transcription regulation to cell cycle checkpoint

Richard Cai, Denise Fischer, Yan Yan-Neale, Hong Xu and Dalia Cohen[1]

Department of Functional Genomics, Novartis Pharmaceuticals Corporation, Summit, NJ 07901, USA

Recent studies have demonstrated that histone deactylases (HDACs) repress transcription by reducing the level of acetylation on core histones and inducing a tight chromatin structure. This compact chromatin structure is prohibitive for transcription factors to gain access to DNA. We and others (Xiao et al 1997, 1999, Sambucetti et al 1999) have demonstrated that transcription of the p21 gene, the inhibitor of cyclin-dependent protein kinases, is under the negative control of HDAC activity. Treatment with HDAC inhibitors significantly enhanced mRNA and protein levels of p21. The induction of p21 is independent of the action of the tumour suppressor gene, p53, a known regulator of p21, indicating a novel mechanism for p21 regulation in the absence of functional p53.

Among the HDACs discovered so far, HDAC1 was the first cloned (Taunton et al 1996) and most extensively studied. HDAC1 and its closely related homologue HDAC2 interact with numerous proteins (Pazin & Kadonaga 1997). The majority of these interactions result in transcription repression following the association of HDAC1 and 2 with proteins bound to promoter regions. In the case of p21, HDAC1 is probably targeted to the promoter through its interaction with DNA-bound Sp1, since a direct interaction between HDAC1 and Sp1 has been discovered (Doetzlhofer et al 1999). Furthermore, the region in the p21 promoter that responded to the inhibition of HDAC1 was found to contain Sp1 binding sites (Sambucetti et al 1999). This likely resulted in reduced acetylation of the core histones around the p21 promoter followed by transcription repression.

[1]This chapter was presented at the symposium by Dalia Cohen, to whom correspondence should be addressed.

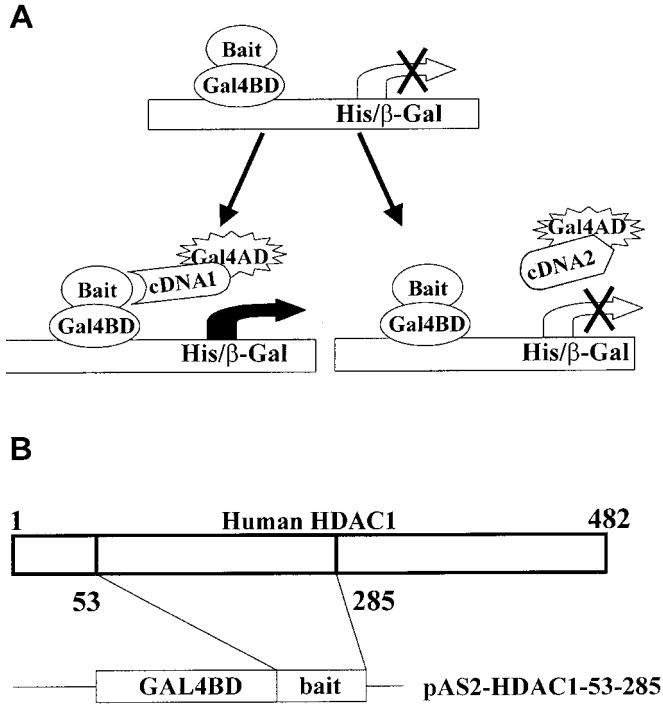


FIG. 1. Yeast two-hybrid screen to identify HDAC1-interacting proteins. (A) The scheme for the yeast two-hybrid screen is described. An interaction between the bait and the polypeptide encoded by cDNA1 brings the activation domain of Gal4p (GAL4AD) to the reporter gene promoter. This leads to the expression of the His and the *β*-galactosidase genes. When the polypeptide encoded by cDNA2 does not interact with the bait, the His and *β*-galactosidase genes remain silent. (B) In our screen, the bait consisted of HDAC1 amino acids 53 to 285. A positive clone, YYN0048, was identified using the *β*-galactosidase assay.

To further study the function and regulation of HDAC1, we searched for novel cellular factors that interacted with HDAC1 by using a yeast two-hybrid screening (Fig. 1A). A large N-terminal region of HDAC1 from amino acids 53–285 out of a total of 482 was used as the bait to search for interacting cellular factors in a HeLa cDNA library (Fig. 1B). A human gene was identified that demonstrated a specific interaction with HDAC1. The gene encoded a polypeptide that was about 30% identical to *Schizosaccharomyces pombe* protein, hus1+p (for hydroxyurea sensitive) and it was therefore named human Hus1, hHus1. The cloning of hHus1 was also reported by Kostrub et al

(1998). Hus1 homologues have also been identified in mouse, *Caenorhabditis elegans* and *Drosophila* (Dean et al 1998, Fig. 2A). *S. pombe* hus1+p was reported to be a checkpoint rad protein that together with five other known rad proteins, relays a signal from DNA damage or replication block to downstream effectors (Kostrub et al 1997, Russell 1998). This results in a G2/M growth arrest in cells suffering DNA damage or replication block (Fig. 2B).

The interaction between HDAC1 and hHus1 was characterised *in vitro* and *in vivo*. *In vitro*, immobilized GST-hHus1 fusion protein bound ^{35}S-labelled HDAC1. When GST-hHus1 binding assays were performed with various HDAC1 deletion mutants, it was found that the HDAC1 region responsible for the *in vitro* interaction was mapped between amino acids 1 and 240. In addition, since the yeast two-hybrid screening indicated that the region between amino acids 53 and 285 of HDAC1 interacted with hHus1, we concluded that the HDAC1 putative region that interacted with hHus1 encompassed amino acids 53 to 240. HDAC1 and hHus1 were also found to interact *in vivo*. In transfected cells, immunoprecipitation of HDAC1-flag precipitated co-expressed HA-hHus1, a flu-epitope tagged hHus1 (Fig. 3), or GFP-hHus1, a green fluorescent protein tagged hHus1 (Cai et al 2000). Furthermore, HDAC1-flag was found to co-immunoprecipitate with rad9 (Cai et al 2000), which is one of the checkpoint rad proteins. The finding that hHus1 interacted with rad1 and rad9 (Kostrub et al 1998, St Onge et al 1999, Volkmer & Karnitz 1999), suggested the existence of a functional complex between HDAC1, hHus1, rad1 and rad 9. This HDAC1–rad 9 interaction might be stabilized by hHus1, which could act as a bridge between HDAC1 and rad9. Taken together, these data indicate that hHus1 is a novel HDAC1 interacting factor.

Our findings that HDAC1 interacts with G2/M checkpoint rad proteins suggested an involvement of HDAC1 in cell cycle regulation. Interestingly, bioinformatics analysis indicated that both hHus1 (Aravind et al 1999) and rad1 (Thelen et al 1999) may contain the so-called PCNA motif that is responsible for the trimerization and binding to DNA of the proliferating cell nuclear antigen (PCNA), a processivity factor for DNA polymerase δ (Gulbis et al 1996). This analysis suggested that checkpoint rad proteins could employ a mechanism similar to that of PCNA binding to DNA. The interaction of Hus1 with

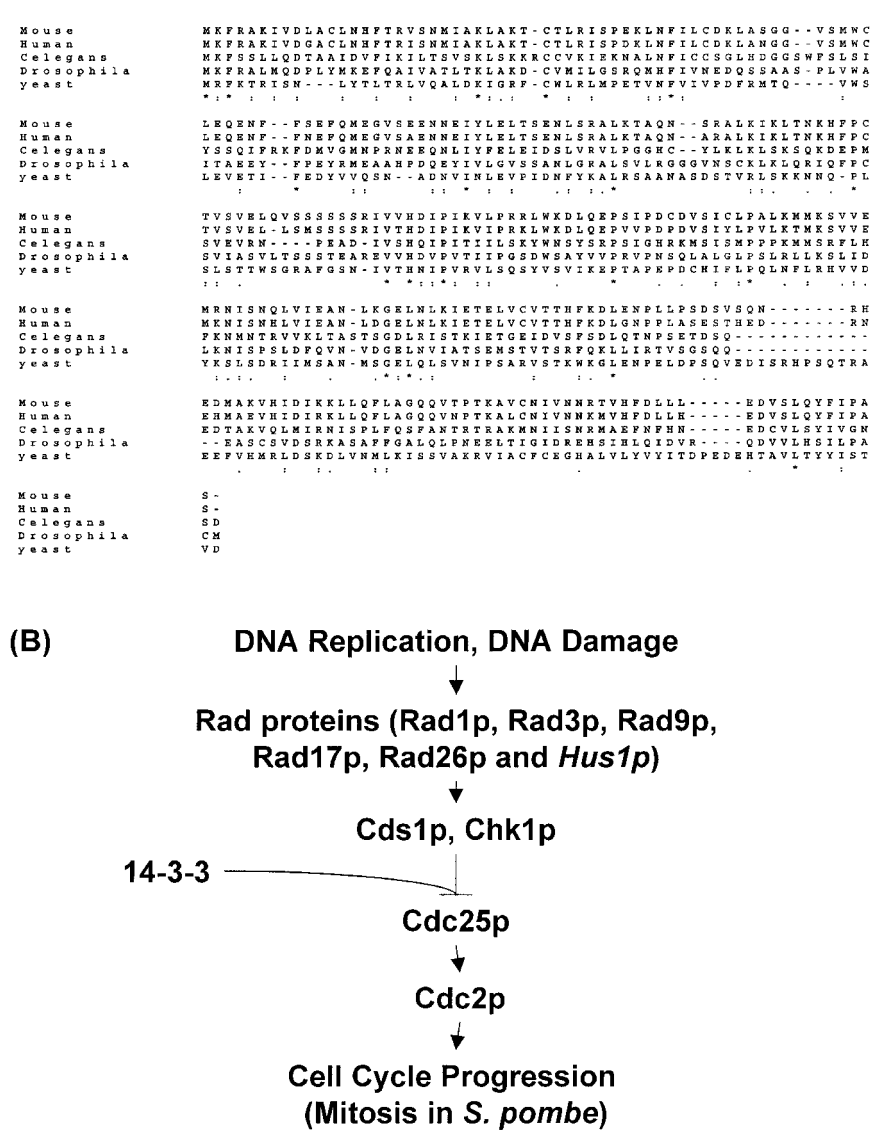


FIG. 2. Hus1 protein is a checkpoint rad protein. (A) Multiple sequence alignment of Hus1 homologues from mouse, human, *C. elegans*, *Drosophila* and *S. pombe*. (B) The *S. pombe* checkpoint regulatory pathway. This checkpoint mediates cell growth arrest at G2/M when cells suffered DNA damage or replication block.

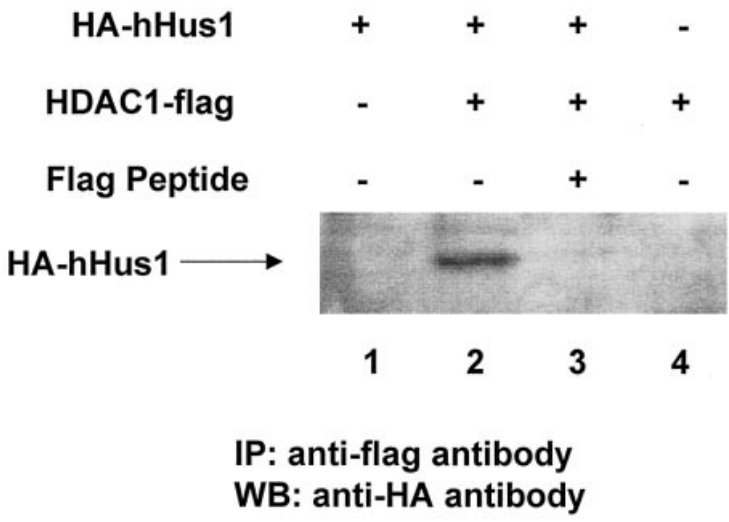


FIG. 3. HDAC1 and hHus1 interact *in vivo*. Expression constructs for either the flag-epitope tagged HDAC1 (HDAC1-flag) or the HA-epitope tagged hHus1 (HA-hHus1) were transfected into COS-7 cells. Immunoprecipitation was carried out with anti-flag antibody. The immunocomplexes were then examined by Western blot using anti-HA antibody.

HDAC1 could lead to chromatin structure modifications that facilitate DNA repair.

References

Aravind L, Walker DR, Koonin EV 1999 Conserved domains in DNA repair proteins and evolution of repair systems. Nucleic Acids Res 27:1223–1242

Cai RL, Yam-Neale Y, Cueto MA, Xu H, Cohen D 2000 HDAC1, a histone deacetylase, forms a complex with Hus1 and Rad9, two G2/M checkpoint rad proteins. J Biol Chem, in press

Dean FB, Lian L, O'Donnell M 1998 cDNA cloning and gene mapping of human homologs for *Schizosaccharomyces pombe rad17*, *rad1*, and *hus1* and cloning of homologs from mouse, *Caenorhabditis elegans*, and *Drosophila melanogaster*. Genomics 54:424–436

Doetzlhofer A, Rotheneder H, Lagger G et al 1999 Histone deacetylase 1 can repress transcription by binding to Sp1. Mol Cell Biol 19:5504–5511

Gulbis JM, Kelman Z, Hurwitz J, O'Donnell M, Kuriyan J 1996 Structure of the C-terminal region of p21(WAF1/CIP1) complexed with human PCNA. Cell 87:297–306

Kostrub CF, al-Khodairy F, Ghazizadeh H, Carr AM, Enoch T 1997 Molecular analysis of *hus1*+, a fission yeast gene required for S-M and DNA damage checkpoints. Mol Gen Genet 254:389–399

Kostrub CF, Knudsen K, Subramani S, Enoch T 1998 Hus1p, a conserved fission yeast checkpoint protein, interacts with Rad1p and is phosphorylated in response to DNA damage. EMBO J 17:2055–2066

Pazin MJ, Kadonaga JT 1997 What's up and down with histone deacetylation and transcription? Cell 89:325–328

Russell P 1998 Checkpoints on the road to mitosis. Trends Biochem Sci 23:399–402
Sambucetti LC, Fischer DD, Zabludoff S et al 1999 Histone deacetylase inhibition selectively alters the activity and expression of cell cycle proteins leading to specific chromatin acetylation and antiproliferative effects. J Biol Chem 274:34940–34947
St Onge RP, Udell CM, Casselman R, Davey S 1999 The human G2 checkpoint control protein hRAD9 is a nuclear phosphoprotein that forms complexes with hRAD1 and hHUS1. Mol Biol Cell 10:1985–1995
Taunton J, Hassig CA, Schreiber SL 1996 A mammalian histone deacetylase related to the yeast transcriptional regulator Rpd3p. Science 272:408–411
Thelen MP, Venclovas C, Fidelis K 1999 A sliding clamp model for the Rad1 family of cell cycle checkpoint proteins. Cell 96:769–770
Volkmer E, Karnitz LM 1999 Human homologs of *Schizosaccharomyces pombe* rad1, hus1, and rad9 form a DNA damage-responsive protein complex. J Biol Chem 274:567–570
Xiao H, Hasegawa T, Miyaishi O, Ohkusu K, Isobe K 1997 Sodium butyrate induces NIH3T3 cells to senescence-like state and enhances promoter activity of p21WAF/CIP1 in p53-independent manner. Biochem Biophys Res Commun 237:457–460
Xiao H, Hasegawa T, Isobe K 1999 Both Sp1 and Sp3 are responsible for p21waf1 promoter activity induced by histone deacetylase inhibitor in NIH3T3 cells. J Cell Biochem 73:291–302

DISCUSSION

Venter: The initial response that you got indicated that this should interfere with all cellular biology, so why doesn't it?

Cohen: There is not just one histone deacetylase, but several members of one family.

Venter: How large is each of these families?

Cohen: The only one we know much about so far is the one I described, HDAC 1, but the family contains 11 members.

Hochstrasser: It makes sense to worry about that. In medicine, when you give Ca^{2+}-blocking agents, for example, they have wide-ranging effects.

Venter: Drugs clearly work despite our knowledge.

Cohen: It is a matter of a therapeutic window. If you can give a drug at an effective dose that doesn't kill the patient, that is OK.

Venter: If you look in yeast, *C. elegans* or *Drosophila* how many different families do you find?

Cohen: Not all the available databases can be searched at the moment.

Venter: How many different ones do you seen in *C. elegans*?

Cohen: Just one. It is interesting that *Saccharomyces cerevisiae* does not have one, but *S. pombe* does.

Fraser: Early on in your paper you showed transcript analysis and proteomics as two key components in your overall functional genomics programme. Do you have any sense as to how often the data you get from

transcript analysis don't agree with changes in levels of protein? Is that something you worry about when you try to integrate data from these two approaches?

Cohen: This is a question which is being asked all the time, and we really need much more data from both approaches to answer it. One thing that we should all be aware of is that we are talking about both closed and open systems. We are only going to see what is on the chip and not what is not on the chip. It is clear that somewhere along the line we will have to use open systems.

Hochstrasser: From what I've seen so far, the correlation between the transcript and the proteomics side is less than 50%. Structural proteins with a long half-life may be abundant while their respective mRNAs have already disappeared. In contrast, secreted proteins may have left the cell while there is still a lot of mRNA for new synthesis within the same cell.

Venter: What about the yeast two-hybrid system? In terms of finding these interactions, how do you sort them out from all the noise and the other data in the background? To go from what typical data is with that system, to come up with this specific interaction would be extraordinarily brilliant work or a chance event. Maybe it is some of both.

Cohen: This effort involved large-scale sequencing. I am sure that there are others here more experienced with the yeast two-hybrid system, and who can streamline this process. A data set for non-specific interactions is available to everyone on a web page.

Mann: I have a question about drug targets. This was partly prompted by a recent workshop (Screens for therapeutic targets and leads: emerging approaches in applied functional genomics, Every, France, June 10–11 1999) where there were people working on interesting and clearly relevant fundamental mechanisms and associated intracellular targets on the one hand (mainly from biotech companies), and people from the pharmaceutical industry who were only interested in working on a very limited set of 'tractable' targets, such as G protein-coupled receptors, ion channels and so on. Someone said that they are not interested in acetylases because they cannot successfully be made into drug targets. How are you going to deal with that problem?

Cohen: This is obviously an important question. The most important approach is to concentrate on disease-specific areas. One of the strengths

in being in a pharma is the integration of the extensive knowledge and the availability of in-house model systems to study disease pathophysiology and the comprehensive *invitro* and *invivo* assays for modelling disease with the new technological approaches contributed by functional genomics.

Hochstrasser: You mentioned gene polymorphism and the way that this can cause differential drug responses in patients depending on their polymorphic variation in certain genes. What about the impact of the environment? As an example, I heard that when you eat certain vegetables, some leaves have been affected by fungi or viruses. As a defence strategy the plant produces salicylate around the lesion, so when you eat the leaves, you may be exposed to a small dose of salicylate. This may cause different backgrounds in the patient from the environment, which may modify the effect of genetic polymorphism.

Venter: I hadn't heard that before, but it is a great example showing why genetic determinism is not absolute. There are increasing data that show polymorphic variation determining the propensity for infectivity of microbial agents. It means that biology isn't all that simple. At least we'll all have jobs for a long time to come!

Lipshutz: There are quite a few other examples of drug interactions with environmental factors. One that has shown up several times is the presence of a bergotamine in the peel of grapefruits that has a major competitive inhibitory effect on cytochrome P450 in the gut and liver, and can change the way that individuals are able to process certain drugs.

Venter: I think all these different issues argue the point we are trying to make in terms of the computational biology: we are going to need phenomenal new computer tools to track and understand what is going on.

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

Mass spectrometry resurrects protein-based approaches in functional genomics

Matthias Mann

Protein Interaction Laboratory (www.pil.sdu.dk), University of Southern Denmark Odense, Campusvej 55, DK-5230 Odense M, Denmark

The anticipated availability of virtually all human gene sequences within the next year will usher in the 'post-genomic era' of biology sooner than expected. We now require large-scale theoretical and experimental approaches which will use the genomic information but add other dimensions of 'functional' information to it. Methods that are already being applied include bioinformatic approaches (including comparative genomics), large-scale two-hybrid screening (currently for small-to-medium genome sizes) and large-scale expression analysis via DNA chip arrays. It would be highly desirable to complement the above approaches with protein-based approaches which would provide information directly at the level of the expressed gene products.

Previously, protein analysis was limited by the lack of sensitivity and throughput of the available methods, such as the Edman degradation. Advances in mass spectrometry over the last few years now make it possible to identify large numbers of gel-separated proteins at minute levels (low femotmole/ low nanogram) (Shevchenko et al 1996, Wilm et al 1996). Stained protein spots are excised from gels, enzymatically digested (usually by trypsin) and the resulting peptides subjected to mass spectrometry. In the matrix-assisted laser desorption/ionization (MALDI) method, the peptide masses are measured with high mass accuracy and the set of masses is then screened against the set of expected tryptic peptide masses for each protein or open reading frame in comprehensive protein databases. Only a few peptide masses are required for unambiguous identification, therefore modified proteins

and protein mixtures can be accommodated as well. This method can be automated and is particularly useful for proteins from organisms with a completely sequenced genome. A throughput of several hundred proteins per day should be feasible with further development of the associated technology (currently under development at Protana and other companies). An alternative method, electrospray tandem mass spectrometry, yields actual amino acid sequence information rather than only a 'mass fingerprint' of the protein. In this method the peptide mixture is sprayed through application of an electric potential to a hypodermic needle through which the liquid is pumped. The liquid then disperses into small, highly charged droplets which evaporate and liberate protonated peptides. Inside the mass spectrometer, these peptides are separated according to mass. After a first selection, a given peptide species is collided with background gas and the resulting peptide fragments separated in another mass separating step (tandem mass spectrometry). The differences in mass between the fragments contain partial information about the amino acid sequence of the peptide. Using a miniaturized version of electrospray (nanoelectrospray) and special database searching algorithms (peptide sequence tags) it is possible to identify proteins on the basis of one or two peptides. Even at the current state of the human genome project, almost all human proteins can be identified via their corresponding expressed sequence tag (EST) entries. These methods have been used with success in a variety of questions involving the identification of low amounts of proteins and also in the 'classical proteomics' approaches (see this volume: Hochstrasser et al 2000, van Oostrum et al 2000).

As shown by our group, the above mass spectrometric methods can also be used to study protein interactions via the analysis of multiprotein complexes. Briefly, proteins of interest can be precipitated using gene tagging or antibody methods, revealing interacting proteins on 1D or 2D gels which can then be identified by mass spectrometry (Lamond & Mann 1997, Neubauer et al 1997). The first example of this technology was the analysis of the yeast U1 snRNP particle, a subunit of the spliceosome. Twenty protein products were identified and later shown to be bona fide members of the U1 subunit. The human spliceosome was similarly analysed by our group, yielding a large number of novel proteins which were identified as EST fragments and

cloned from that information. Co-localization of the novel factors with known splicing factors was used as a rapid method of initial verification. More than 30 protein complexes have now been analysed by these methods, and we have shown that this technology can be scaled up to large numbers. Significant biological results have already been obtained both in structural protein complexes and in transient complexes such as the ones involved in signalling (Neubauer et al 1998, Yaron et al 1998). In principle this technology can lead to a protein interaction map of the cell. The approach should be accompanied by bioinformatics tools which interpret the empirically found interactions. We conclude that mass spectrometry of multiprotein complexes is a valid approach which rapidly yields functional information on open reading frames identified in sequencing projects.

References

Hochstrasser D, Sanchez J-C, Binz P-A, Bienvenut W, Appel RD 2000 A clinical molecular scanner to study human proteome complexity. In: From genome to therapy: integrating new technologies with drug development. Wiley, Chichester (Novartis Found Symp 229) p 33–40

Lamond AI, Mann M 1997 Cell biology and the genome projects—a concerted strategy for characterizing multi-protein complexes using mass spectrometry. Trends Cell Biol 7:139–142

Neubauer G, Gottschalk A, Fabrizio P, Séraphin B, Lührmann R, Mann M 1997 Identification of the proteins of the yeast U1 small nuclear ribonucleoprotein complex by mass spectrometry. Proc Natl Acad Sci USA 94:385–390

Neubauer G, King A, Rappsilber J et al 1998 Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. Nat Genet 20:46–50

Shevchenko A, Jensen ON, Podtelejnikov AV et al 1996 Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. Proc Natl Acad Sci USA 93:14440–14445

van Oostrum J, Mueller D, Schindler P 2000 From proteomics to functional analysis. In: From genome to therapy: integrating new technologies with drug development. Wiley, Chichester (Novartis Found Symp 229) p 41–53

Wilm M, Shevchenko A, Houthaeve T et al 1996 Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. Nature 379:466–469

Yaron A, Hatzubai A, Davis M et al 1998 Identification of the receptor component of the IkappaBalpha-ubiquitin ligase. Nature 396:590–594

DISCUSSION

Venter: My understanding was that the limiting step in this process is the initial 2D gel separation. When Roche tried to do this with *Haemophilus*, I thought that they could only get sequences of roughly 40% of the proteins off the gel. I don't know what the yeast data are like.

Mann: There are different approaches for separating the proteins. In our approach, we virtually always employ a first affinity purification step, which is then followed by 1D or 2D electrophoresis. Particularly when we are looking for new receptors, we cannot use 2D gels because the proteins are often 100–200 kDa in size, so they wouldn't show up in the 2D gel in the first place.

Venter: With 1D separation, you can hide a lot of data. How do you separate out artefacts?

Mann: We were previously very concerned about precipitating proteins which were artefacts. However, with a good purification and good controls we see surprisingly little contamination. In the case of the human spliceosome (Neubauer et al 1998) we only found four contaminations out of more than 70 sequenced protein spots. These were readily apparent as such; all the other proteins—so far as we know—are genuine members of the human splicesome.

Hochstrasser: I would add that the affinity purification of proteins prior to the 2D gel analysis is an essential step. Without doing this, the use of 2D gels doesn't make much sense. Pre-fractionation of a sample is very useful. 15% of expected proteins from the genome don't show up on 2D gels because of their high hydrophobicity. 1D gels should therefore be used to separate those hydrophobic proteins.

Goodfellow: Every technique has its limitations. The limitation here is that we do not know how to define a biologically meaningful interaction in terms of affinity between two proteins. This isn't a criticism of your technique, but you could be missing 50% of the important interactions which occur in those complexes, because they are weak interactions.

Mann: We don't claim that our technique is exhaustive.

van Oostrum: It may become a little more transparent if the immunoprecipitation is followed by washes at different stringencies. Then you will obtain additional information about the affinities of the components in the complex.

Goodfellow: I'm more concerned about the question of what is the affinity of a biologically meaningful interaction.

Mann: It is an important issue. Sometimes, for instance, three proteins will come together for a very short time and none of the pairwise affinities will be very high.

Venter: Lesson number one, when people first start doing searches of any protein against the databases, is that if you go down far enough on any matrix you can start finding 10mers, 12mers, 14mers that are exact matches. I got confused with this very early on with the β receptor where I found a 15 amino acid stretch that matched ragweed pollen exactly. It was such a lovely association that we thought it must be meaningful. So we must be careful of these false-positive results.

Mann: We have identified thousands of proteins and have not run into this problem. Either we get a very clear match with a peptide mass fingerprint or we go on to sequence the protein. When we sequence it we will have several peptides which each uniquely identify the protein or open reading frame in question.

Venter: You can have multiple hits.

Mann: Then you will know that you have multiple hits, but it's different from homology, because any amino acid change that leads to a molecular weight change will be picked up. It is not often that you have two tryptic peptides that match the same protein by chance, unless they are in the same gene family.

Venter: But your experiments are degenerate. You are going from the peptide code back to the DNA sequence, and so you have to look at the multiple possibilities.

Mann: If you have a tryptic peptide that is 12 amino acids, and you know the cleavage site, too, you in effect have 13 amino acids. If you look in the whole database, how many times will two such peptides randomly occur in another protein? This would be very rare indeed, and at least in the case of completely sequenced genomes you would still get both possibilities listed. The degeneracy of going from the protein data to search the DNA sequences is not as great as is first appears. Even in the case of reading frame errors in DNA sequencing, the peptides all have to occur in the same direction and a given stretch of DNA can only code for three different peptides. So this does not add appreciably to the statistics.

Venter: There are other techniques with the AFLP analysis that just use a few base pairs on either side of a restriction site that work frequently.

Mann: In any case, typically our data cover at least 60–80 amino acids, even if we initially use much fewer for searching the genomic databases, all these data have to fit the final open reading frame, so in practice there is absolutely no uncertainty about the identification of the gene. We

routinely search a protein database of more than 300 000 entries, so the statistics will not get worse when all human genes are known. Apart from 'deconvoluting' very complex protein mixtures, the only difficulty is in distinguishing two nearly identical forms of a protein. In this case one is dependent on sequencing the peptide or the peptides that contain the difference between the two genes

Hochstrasser: It's a very good question. When a genome is known, how many amino acids from either end of a protein do you need to know to be able to identify a protein with certainty? In *Escherichia coli*, you need to know just three amino acids from the C-terminus and four from the N-terminus. In yeast, you need to know six amino acids. In eukaryotes it is obviously higher.

Venter: As soon as you get out of a microbial genome the artefacts go up exponentially, so you need more. This is the concern.

Reference

Neubauer G, King A, Rappsilber J et al 1998 Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. Nat Genet 20:46–50

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

A clinical molecular scanner to study human proteome complexity

Denis Hochstrasser*, Jean-Charles Sanchez*, Pierre-Alain Binz*†, Willy Bienvenut* and Ron D. Appel†

**Central Clinical Chemistry Laboratory and †Swiss Institute of Bioinformatics, Geneva University Hospital, CH-1211 Geneva 14, Switzerland*

Diagnostic means capable of recognition. Prognostic means knowing in advance.

The process to establish a diagnosis and to evaluate the prognosis of identified diseases is essential for selecting and evaluating an appropriate treatment or testing new drugs.

Every patient is unique. With the exception of identical twins, each human being has a unique genetic background and is submitted to various external influences modifying gene expression. The final patient phenotypes are therefore extremely complex and unique. Grouping patients into categories by establishing a diagnosis should always be challenged. At least, the most sensitive and refined procedures should be used to identify and classify patient diseases into sub-categories. Tools should be developed to highlight efficiently relevant molecular pathways that are involved in defined disease processes. They should measure the effect of each phenotype on the disease process and the respond to treatment.

As mentioned above, every patient has a unique genetic predisposition and is under major external influences (Fig. 1) (Hochstrasser 1998). The genetic predisposition can be analysed on a large scale by DNA array or other high throughput DNA tests, although many DNA changes may not be relevant for the disease process studied. The environmental influences should also be studied at the expression level, i.e. the mRNA and/or the protein level. Massive mRNA studies are performed using for example reverse transcriptase PCR and DNA chip technology (Wang et al 1999).

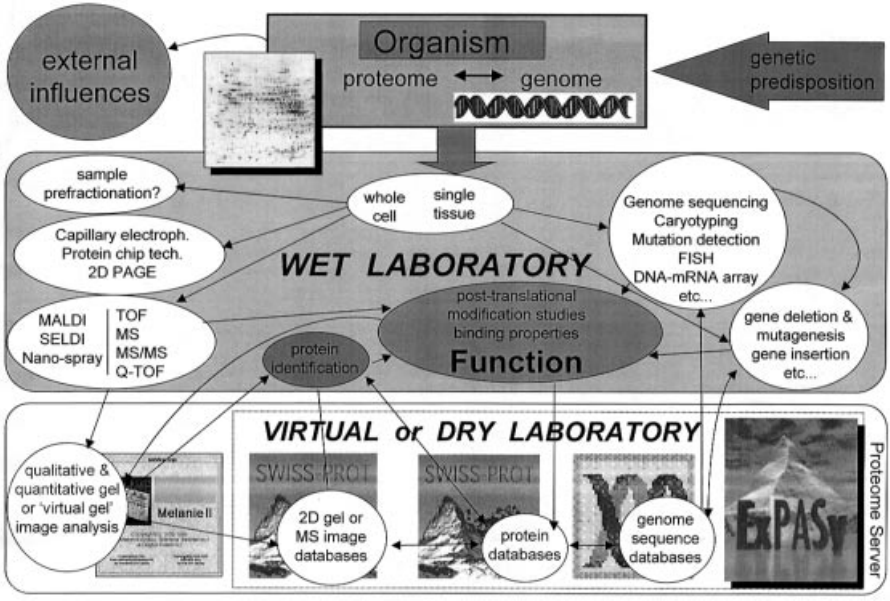


FIG. 1. Scheme showing the relationship between the real (or 'wet') laboratory and the virtual (or 'dry') computer laboratory. The central question of genomics and proteomics is to find the function of genes and proteins. The core piece of proteomics is curated and annotated databases such as SwissProt. (Modified from Hochstrasser 1998.)

However, the correlation between the level of mRNA and the protein concentration is often weak or even absent in some cases (Anderson & Seilhamer 1997).

In addition, most if not all human proteins are post-translationally modified and thus one gene expresses between three and more than 20 final proteins with unique 3D structures (Wilkins et al 1996). Consequently, the proteins expressed by a genome, the proteome, should also be analysed and quantified. Only proteome studies will highlight the functional products of many genes and will underline the epigenetic network regulating cell or tissue function (Strohman 1994, 1995, 1997).

Assuming that the human genome contains 50 000–100 000 genes, the global human proteome should most likely display half a million post-translationally modified proteins or products. With the technology available today only a very small fraction of the human proteins can be

TABLE 1 Comparison between nucleic acids and proteins

	Nucleic acids	*Proteins/peptides*
Belongs to	Genomes	Proteomes
Level	Information	Product
Number of building blocks	4+	$\geqslant 20$
Solubility in water	High	Highly variable, sometimes very low
Prediction of behaviour	Easy	Difficult
Number of specific cleavage enzymes	Very many (>300)	Few (<12)
Possible propagation or amplification	Easy	No (except prions?)
Sequencing speed	Very fast	So far, very slow

analysed readily. This complexity requires some type of pre-fractionation or purification of the sample and enrichment of the proteins of interest. The task is difficult not only because of this large number of proteins, but also because of their tremendous chemical heterogeneity, their behaviour and the large dynamic range of their concentration.

Differences in pI, size, hydrophobicity and half-life of proteins is tremendous. The difference between the least and most concentrated protein exceeds 12 logs. In contrast to DNA analysis, no simple protein amplification process exists (Table 1).

Several possibilities or technology pathways are being explored to partially resolve this problem of complexity. Each protein form has a unique charge or pI under a defined physicochemical condition, a precise mass, a unique fine 3D structure and related binding properties or function, a unique amino acid sequence, and carries a specific set of post-translational modifications. Each of those characteristics can be highlighted alone. But only a combination of several of them allows a proper and unique identification and partial characterization of the protein when the genome of the related organism is known. For example, the combination of protein affinity chips and a matrix-assisted laser desorption/ionization (MALDI) mass spectroscopy (MS) scanner will allow the precise identification and partial characterization of thousands of proteins in a very short time. The combination of isoelectric focusing (IEF) capillary electrophoresis and MS will

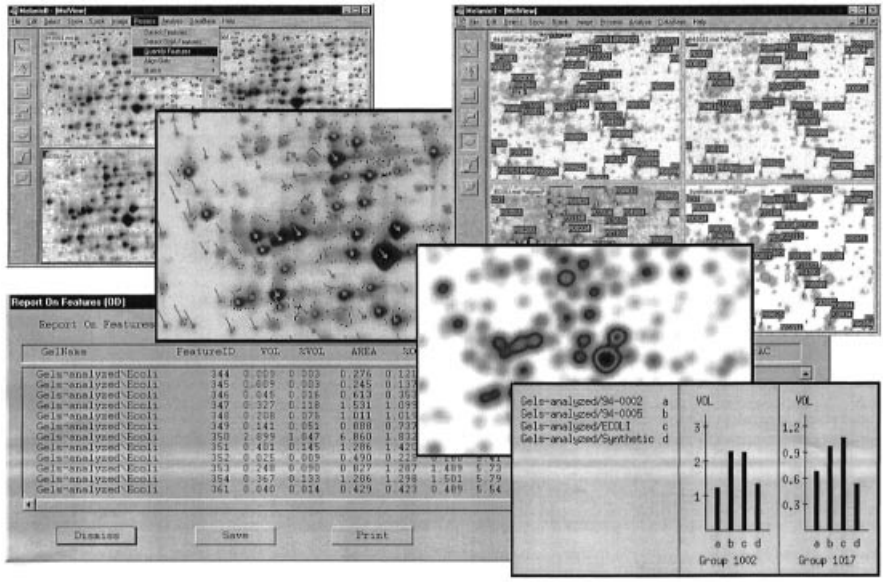


FIG. 2. Multiple screens displayed by the MELANIE software and showing, from left to right, 2D PAGE spot detection, automatic gel matching with matching vectors, protein spot quantitation and their identification with pink labels though internet connection to Swiss 2D PAGE database.

highlight characteristic spectra and will allow the identification of many proteins in a known sample. But today, the orthogonal separation of proteins by charge and by size followed by several types of protein fragmentation and precise mass analysis of the fragment is, in our opinion, one of the most efficient and reliable proteomic approaches (Aebersold 1993, Mann & Wilm 1994, 1995).

The other is to study protein binding properties to discover new ligands and their function.

The interpretation of data obtained by the combination of these techniques and also by DNA analysis requires powerful bioinformatic tools, such as MELANIE (Fig. 2) (Hochstrasser et al 1995, Wilkins et al 1996, 1997, 1998) and intranet access to non-redundant, curated and annotated databases (Bairoch & Apweiler 1997, Bairoch et al 1997, Hawkins et al 1999). We combined reproducible 2D PAGE techniques using narrow immobilized pH gradient (IPG) precasted gel strips, precasted mini SDS PAGE gels, a new digesting-transblot procedure

and direct MALDI scanning of a collecting PVDF membrane to display bacterial and human proteome sections (Hochstrasser 1998). Preliminary results indicated that the full genome of *Eschericia coli* could be displayed by such methodologies and, with minor technology progress, thousands of proteins could be detected and partially characterized within a few days.

Future refinement and miniaturization of such methodologies should further increase the sample throughput and provide a molecular scanner for clinical applications (Hochstrasser et al 1991). Indeed the precise identification, quantitation and partial characterisation of thousands of proteins in tissue biopsies or fluids should improve the diagnostic processes and disease prognostic evaluation.

Acknowledgements

This work was supported by the Swiss National Fund for Scientific Research, the Helmut Horton and the Montus Foundation.

References

Aebersold R 1993 Mass spectrometry of proteins and peptides in biotechnology. Curr Opin Biotechnol 4:412–419

Anderson L, Seilhamer J 1997 A comparison of selected mRNA and protein abundances in human liver. Electrophoresis 18:533–537

Bairoch A, Apweiler R 1997 The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acids Res 25:31–36

Bairoch A, Bucher P, Hofmann K 1997 The PROSITE database, its status in 1997. Nucleic Acids Res 25:217–221

Hawkins V, Doll D, Bumgarner R et al 1999 PEDB: the Prostate Expression Database. Nucleic Acids Res 27:204–208

Hochstrasser DF 1998 Proteome in perspective. Clin Chem Lab Med 36:825–836

Hochstrasser DF, Appel RD, Vargas R et al 1991 A clinical molecular scanner: the Melanie project. MD Comput 8:85–91

Hochstrasser DF, Appel RD, Golaz O, Pasquali C, Sanchez JC, Bairoch A 1995 Sharing of worldwide spread knowledge using hypermedia facilities & fast communication protocols (Mosaic and world wide web): the example of ExPASy. Methods Inf Med 34:75–78

Mann M, Wilm M 1994 Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem 66:4390–4399

Mann M , Wilm M 1995 Electrospray mass spectrometry for protein characterization. Trends Biochem Sci 20: 219–224

Strohman RC 1994 Epigenesis: the missing beat in biotechnology? Biotechnology (NY) 12:156–164 (erratum 1994 Biotechnology 12:329)

Strohman RC 1995 Linear genetics, non-linear epigenetics: complementary approaches to understanding complex diseases. Integr Physiol Behav Sci 30:273–282

Strohman RC 1997 The coming Kuhnian revolution in biology. Nat Biotechnol 15:194–200

Wang K, Gan L, Jeffery E et al 1999 Monitoring gene expression profile changes in ovarian carcinomas using cDNA microarray. Gene 229:101–108

Wilkins MR, Sanchez JC, Williams KL, Hochstrasser DF 1996 Current challenges and future applications for protein maps and post-translational vector maps in proteome projects. Electrophoresis 17:830–838

Wilkins MR, Lindskog I, Gasteiger E et al 1997 Detailed peptide characterization using PEPTIDEMASS—a world-wide-web- accessible tool. Electrophoresis 18:403–408

Wilkins MR, Gasteiger E, Tonella L et al 1998 Protein identification with N and C-terminal sequence tags in proteome projects. J Mol Biol 278:599–608

DISCUSSION

Goodfellow: When you're analysing a 2D gel and you observe a large number of variants in a particular protein, can you define the cause of the variation? Can you make a difference map which tells you the differences between these proteins?

Hochstrasser: There are many answers to your question. Often when you see a protein moving to the left and going upwards on the 2D-PAGE image, it is often sialic acid that is responsible. If a protein moves to the left in steps with no obvious upwards shift, it is often due to phosphorylation.

We have developed bioinformatic tools that will help identify the differences from the mass spectra. One of these, which will be published on the web soon, is called glycomod. It helps to identify some sugar modifications. We already have another tool, called findmod, which helps to discover 21 potential modifications on the protein. Unfortunately, not all peptides fly: for example, phosphopeptides do not fly well in a mass spectrometer. Other more time-consuming techniques must then be used.

Goodfellow: This is what I was getting at. If I were to make a list of the different modifications which are known to occur on proteins, how many of those can we currently recognize?

Hochstrasser: Our bioinformatic tools help to identify 21 of them, but this is not a lot when we think of the modifications that occur in nature.

Goodfellow: Is that 21 out of 22, or 21 out of 50?

Mann: That is not the right way to phrase the question. When you have a spot on a gel, you need two peptides or so to identify it. If you want to know the complete primary structure of a protein, you're asking 100% sequence coverage, that is, you need to mass measure or sequence *all* the peptides. In the case of non-stochiometic phosphorylation it's even worse. You may have sequenced the non-phosphorylated form of the

peptide but overlooked the 3% of a phosphorylated version of the same peptide (Neubauer & Mann 1999). Looking for all these modifications, even all the unknown ones, is definitely possible but it's a whole different ballgame in terms of time and starting material required.

Goodfellow: But the biological interest of those proteins lies in why they're different. That is why I'm pushing this issue.

Mann: I agree completely, and mass spectrometry is the only generic technique to get at the differences between these alternative forms of the same gene product. Nevertheless, analysing all post-translational modifications is another world from the one we are entering with high throughput protein identification techniques now. As an example, in the biotech industry, it can take six months to exhaustively characterize a protein intended as a therapeutic. And here you even have large amounts. There are also still many unknown or unusual modifications. Using the peptide sequence tag approach (Mann & Wilm 1994), we can match peptides to databases even in the presence of 'errors'—that is, discrepancies between the sequence in the database and the peptide being measured. We have a list of unexplained mass differences which could represent novel chemical or biologically relevant modifications but nobody has the resources to look into all these questions.

Venter: But even if you take your number of 21 known modifications, if you get those in various combinations, how do you sort them out? Do either of you have examples where you have clearly identified splice variants on top of all these post-translational modifications?

Mann: The issue is one of sensitivity and sequence coverage. For complete coverage you need to sequence every last peptide, so you need a lot more material than if you are just identifying a protein. Generic methods are emerging for the most common and currently most interesting types of modifications, such as phosphorylation and glycosylation. We have tried to use affinity-based approaches in another set of experiments rather than try to get the complete primary structure from one spot on the original gel. In this way, one should be able to generate much more material for study.

Goodfellow: I was aware of the problem with phosphopeptides. Are there modifications which make it difficult to analyse peptides?

Mann: Yes. Glycosylation in which the sugar part is very large, labile and heterogeneous is difficult to analyse. Similarly, other modifications which are large or chemically behave very differently to the peptide to which they are attached can be difficult.

Goodfellow: And lipid modifications?

Hochstrasser: This is something that we are currently studying. You can make lipids fly in the mass spectrometer.

Mann: This is all do-able, it is just an issue of sensitivity.

References

Mann M, Wilm MS 1994 Error tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem 66:4390–4399

Neubauer G, Mann M 1999 Mapping of phosphorylation sites of gel-isolated proteins by nanoelectrospray tandem mass spectrometry: potentials and limitations. Anal Chem 71: 235–242

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

From proteomics to functional analysis

Jan van Oostrum, Dieter Mueller and Patrick Schindler

Novartis Pharma AG, Functional Genomics Area, Protein Sciences Unit, CH-4002 Basel, Switzerland

With the advances made in sequencing genomes, proteomics—the study of the expressed part of the genome—has become a major technology in the field of functional genomics. Changes in the expression and structure of most cellular proteins can be displayed and identified using two steps of protein separation. Two-dimensional polyacrylamide gel electrophoresis (2D PAGE) is still the only method available which is capable of simultaneously separating thousands of proteins. The first dimension of 2D PAGE is isoelectric focusing, during which proteins are separated in a pH gradient until they reach the pH of the stationary phase where their net charge is zero, also referred to as the isoelectric point (pI) of the protein. In the second dimension the proteins are orthogonally separated further by electrophoresis in the presence of sodium dodecyl sulfate (SDS PAGE) based on their molecular mass. Standard 2D gels covering in the first dimension a pH gradient from 3–10 allow routine separation of about 3000 proteins. However, by the use of a series of 2D gels, each covering only 1–2 pH units (ultrazoom gels), about 20 000 protein species can be visualized. Such systems will, for example, allow the visualization of a 'near complete' proteome of *Drosophila*.

Protein identification is nowadays based on mass spectrometric analysis of enzymatic hydrolysates prepared by in-gel digestion. Mostly, characteristic fingerprints of peptide masses obtained by matrix-assisted laser desorption/ionization (MALDI) mass spectrometry and/or partial sequences of selected peptides determined by nanoelectrospray tandem mass spectrometry are matched against databases, including expressed sequence tag (EST) data. The methods for interfacing these high-sensitivity 'downstream' analytical techniques to gel electrophoresis have matured and the complete approach can now be applied reliably at

the femtomole level of gel-isolated proteins. Protein identification is, however, only a first step in protein characterization and detailed analysis has to consider post-translational modifications (PTMs) as a major factor influencing structure and function. The detection of phosphorylation, as the most abundant PTM already at the first level of identification using high-throughput MALDI mass spectrometry, is of prime importance.

In differential proteome studies of Taxol-treated versus untreated human 697 cells, seven distinct protein spots within the pI range 5–6 and apparent molecular mass range 18–23 kDa were identified as stathmin (Fig. 1) with spots A–F being up-regulated. Differences in PTMs could be the reason for the different electrophoretic mobility and phosphorylation is an obvious option, since stathmin contains four known phosphorylation sites at Ser15, -24, -37 and -62 (Fig. 1). To separate potential phosphorylation sites and to achieve optimal sequence coverage, we successfully used a double enzymatic strategy with trypsin and parallel Glu-C treatment together with reflector MALDI detection of phosphopeptides. In MALDI, Ser and Thr phosphorylated peptides form labile protonated ions, which even after the acceleration step lose phosphoric acid (Fig. 2). The kinetic energy of the product ions is therefore reduced by the energy imparted to H_3PO_4. As a consequence these ions penetrate the reflector less than those with full kinetic energy and their focusing is bad. The resulting reduced resolution of the product ions and their reduced mass difference to the phosphorylated parent (about 94 instead of 98 Da), which is illustrated for phosphopeptide 27–42 derived from protein spot B in Fig. 2 (*cf.* m/z 1783.9 and parent m/z 1877.96), allows a convenient assignment of phosphopeptides. Together with nanoelectrospray partial sequencing, the prevailing phosphorylation status of all seven stathmin isoforms was determined. Spot G corresponds to the unphosphorylated species, spot F mainly to Ser37 monophosphate and spot E to Ser24, -37 diphosphate. Spots D and C are isomeric triphosphates (Ser24, -37, -62 and Ser15, -24, -37) and the two remaining spots A and B are phosphorylated at all four sites (Ser15, -24, -37 and -62). They are probably derived from the previously reported two isoforms of stathmin, differing by a yet unknown modification. Since the major isoforms differ mainly in the extent of phosphorylation, most of the information obtained by 2D PAGE is retained in a simpler separation

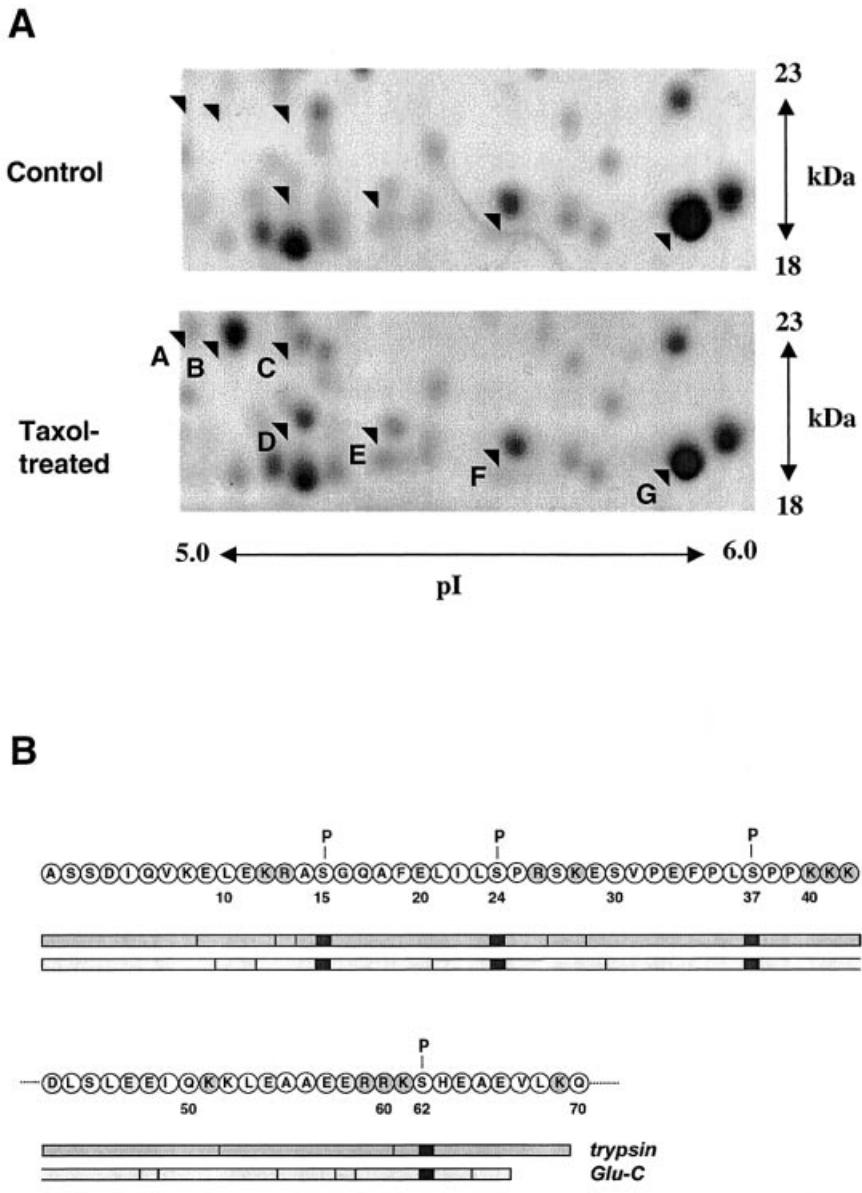


FIG. 1. (A) Detection of differentially expressed proteins by 2D PAGE (stathmin range). (B) Stathmin sequence with enzymatic cleavage sites.

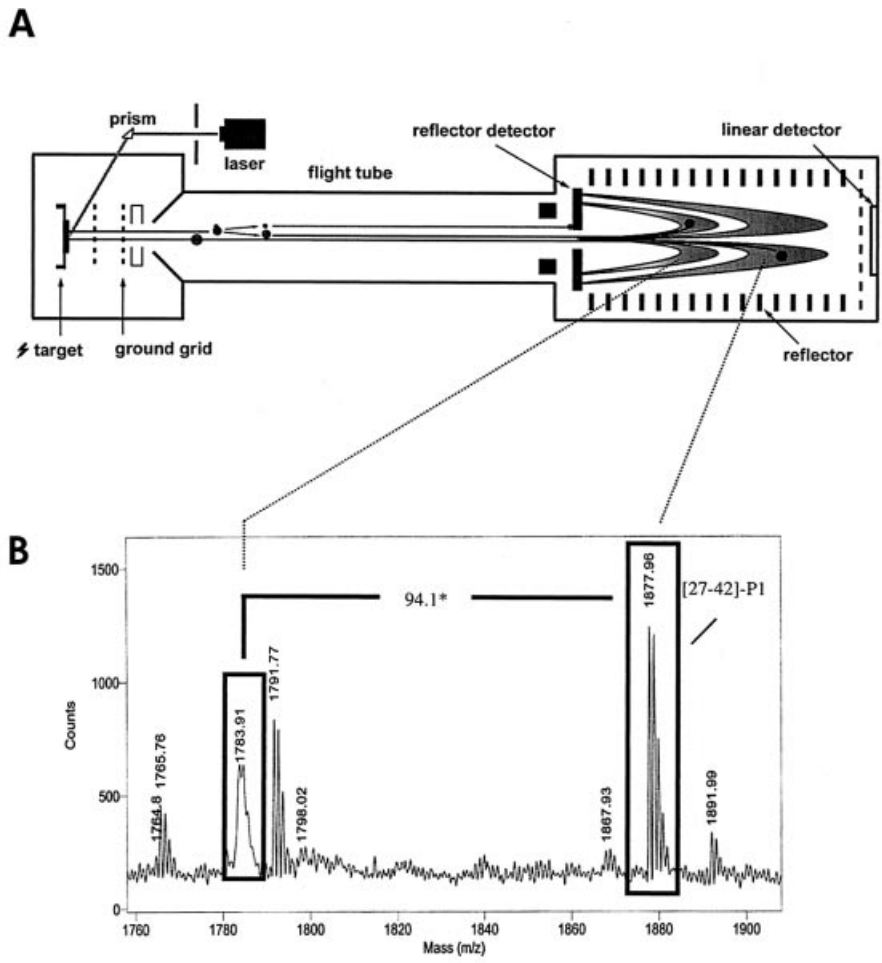


FIG. 2. (A) Identification of phosphopeptides by reflector MALDI-TOF mass spectrometry. (B) Partial reflector MALDI MS derived from protein spot B.

system using native 1D electrophoresis, Western blotting and anti-stathmin antibody detection. This system allows us, for example, to correlate different isoforms with the action of Taxol in time-course experiments. In this way a good correlation between G2/M arrest and appearance of tri- and tetra-phosphorylated species after Taxol treatment of SW-2 cells was observed. In these species, positions 15 and/or 62 are additionally phosphorylated. Interestingly, phosphorylated isoforms are also known to be involved in the regulation of microtubule dynamics. To

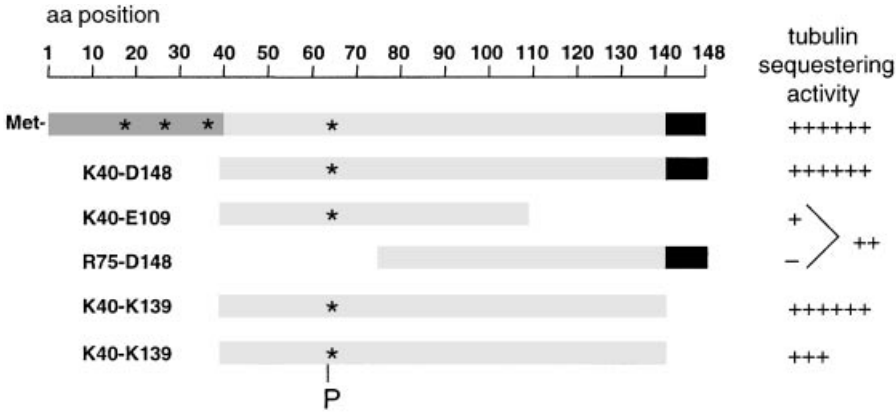


FIG. 3. Stathmin constructs and corresponding tubulin sequestering activity. Stars mark the phosphorylation sites 15, 24, 37 and 62. First line corresponds to wild-type plus N-terminal methionine.

detect whether Ser15 and Ser62 take part in the formation of a potential tubulin/stathmin complex, we prepared several specific stathmin constructs and tested these for tubulin sequestering activity (Fig. 3). Removal of the 39 N-terminal and the nine C-terminal amino acids has no influence, while residues 40–139 are essential for tubulin sequestering. The individual regions 40–109 and 75–148 are practically inactive, but a mixture of both reconstitutes part of the original activity. Ser62 is therefore the only phosphorylation site involved in direct binding to tubulin. Its *in vitro* phosphorylation in the 40–139 construct decreases the sequestering activity by a factor of about two and thus modulates the binding activity of stathmin. For the further localization of the interacting area of stathmin and tubulin, the synthetic peptide 54–72 was added in a 100-fold excess to the complex in a competitive inhibition experiment. Stathmin activity was completely suppressed and tubulin depolymerization was as slow as in its absence. The region around Ser62 is therefore essential for stathmin–tubulin binding and represents a potential target for the interaction of drug candidates.

Conclusion

The direct assignment of Ser/Thr phosphorylated peptides in reflector MALDI mass maps is feasible even in complex mixtures. A double

enzymatic strategy employing trypsin and Glu-C in parallel was essential to separate potential phosphorylation sites of isoforms. A major advantage of the method presented is that knowledge about peptide phosphorylation at this stage already allows partial assignment of modified sites.

On the basis of this fast phosphopeptide characterization, subsequent experiments established a relation between Taxol action and reduced stathmin–tubulin complex formation by phosphorylation of Ser62. The binding site was mapped to a region surrounding Ser62 by selected constructs and by competitive action of peptide 54–72. In summary, this study demonstrates that the step from proteomics to functional analysis of a potential drug target is quite feasible.

DISCUSSION

Rubin: If you can only see proteins that are more than 400 copies per cell, and there are rarely more than 1000 of those in a human cell, you should be able to see all the human proteins that are of sufficient abundance. Is that the case?

van Oostrum: I think so. The detection sensitivity allows you to detect everything above 400 copies per cell, but we probably don't see all of them using 2D gels.

Venter: But that was with intensive radiolabelling.

Rubin: If you want to improve the number of proteins detected, the issue is therefore not getting better resolution of gels, it is getting better sensitivity of detection.

Hochstrasser: It is not just a question of sensitivity; it is also one of separating the proteins. In humans, the difference between the most abundant protein and the least is in the range of 10^{12} orders of magnitude. Using silver staining methods, the difference in concentration between the darker spots and the fainter ones does not exceed 10^4. We are missing 10^8 or we need a lot more on the gels, and this is why you need to spread the proteins on the gel to be able to see all of them.

Venter: Coming back to the paper by Bill Efcavitch (Efcavitch 2000, this volume), it is interesting that although there is elegant technology for detecting the peptide sequence, we are still dealing with gel separation of proteins. It is being refined in pH range, but many of the

advances in DNA sequencing have been getting away from gels into new matrices. The 2D gel seems to be the unchanging tool in this area.

Mann: If you want to see all the proteins, it is not only the sensitivity that is limiting, but also the contrast between the large proteins and small ones, and the hydrophobic ones. These get lost in the 2D gel. It is important to realize what you want to address. For some situations, 1D gels are actually better, such as when the protein complexity has been reduced to a number compatible with the limited resolution of 1D gels of up to 100 proteins. In that case you are reasonably sure you will not lose your protein. In many instances, such as the receptor molecules above 100 kDa that I mentioned in my paper (Mann 2000, this volume), 1D gels are the only means of visualizing the proteins in any case.

van Oostrum: Our efforts are ongoing to move away from 2D gels. I don't think we are anywhere near being able to develop systems as good as those used in DNA sequencing, but there are serious attempts to develop non-gel based multiseparation systems. We face some problems using microsystems with the amount of sample that is required to start out with to obtain a separation of many thousands of proteins and still have enough of each of those proteins for analysis with mass spectrometry.

Efcavitch: The discriminatory power that is being talked about here for proteins is much greater than is current in DNA sequencing. It may be troublesome to handle these gels, but they do give incredible resolution.

You are faced with a choice between looking at global changes versus trying to look at very discrete changes: I think that's what you were hinting at. You can't do both in one gel. Instead, you make a decision to look at either the global pattern of change or to home in on specific modifications.

Mann: One basic question is whether you're looking for expression differences or protein–protein interactions, or the complete native structure of a given gene product.

Venter: There is a disconnect that I'm concerned about. You talk about this resolution of getting every protein over 400 copies per cell, yet as far as I know that hasn't happened with even a single microbial genome yet, in terms of being able to get all the genes expressed and to show up on a 2D gel. Is there are a limited resolution and the spots are buried underneath one another? Or is it actually an expression problem?

van Oostrum: Using the approaches I described, making use of overlapping zoom gels, you get 80% of all proteins on your gels.

Hochstrasser: I would say that you can see 85% on the gel. 15% won't be seen, at least for the time being, because of the hydrophobicity problem.

Venter: There is a study from Hoffmann-La Roche on *Haemophilus* where only they found 40% of the proteins (Fountoulakis et al 1997). Is this because they didn't look with high enough resolution?

van Oostrum: The figure I quoted of 80% is from very recent work using a series of overlapping 2D gels, each with a 1.5 pH interval, covering the pI range from 3.5–9.5.

Hochstrasser: Hoffmann-La Roche took a particular approach. The gels were sliced in small pieces. But the pieces were still a few millimetres wide and contained several proteins. If their resolution was better, such as that obtained with zoom gels, then the spots would be spread out and they could pick them one by one. Then in the mass spectrometer, they would not have a gemisch of proteins and peptides, simplifying data analysis.

Venter: How scaleable is the zoom gel technology?

Mann: It is possible to do this, but then the problem is that you have 10 times as many gels to analyse.

Venter: What's the cost of this?

Hochstrasser: If you are trying to identify the gene products, it is cheap. The first dimensional gel costs about US$10, and the second dimension mini SDS-PAGE precast gel costs also around $10. But you can keep a mass spec guy busy for a few months looking at all of the data from the final 2D PAGE with 1000 spots.

van Oostrum: The main expense is the cost of the people doing the work. A proteomics facility requires 15–20 people.

Hoffman: How much material do you need? In malaria, the major target for a chemoprophylactic drug or vaccine is the early liver stage of the parasite life cycle. We plate 100 000 human hepatocytes *in vitro* and infect them with 100 000 sporozoites. Our yield is only about 30 infected hepatocytes in a sea of 100 000 cells. Is there any way of comparing infected hepatocytes with the non-infected ones using this type of technology? Or, if you were to do laser-guided dissection of single cells, how many would you need in order to use this technique?

van Oostrum: If you used very sensitive radiolabelling you could probably do full scale analysis on the level of 10 000 spots with 10–50 000 cells.

Goodfellow: When in the future do you reach the point where microarray-type approaches invert the problem? It might be possible to use phage-display libraries to make an antibody for every protein (and variant) produced by the genome. Theoretically, one might be able to start with a sequence, express that, pull out the antibody and then put 100 000 antibodies on a grid. The problem becomes completely different. Is this too fanciful?

Lipshutz: If you could make the library, you could sort it on an array.

Goodfellow: We have the tools to be able to select for making phage antibodies. As you can select for one, there is no reason why you can't think of selecting for a large number at once.

van Oostrum: But what you are then doing is switching from proteomics, which is a very open system, to a closed system. I'm not sure that we should do this.

Goodfellow: You have already accepted the fact that you are only assaying 80% of the peptides.

Mann: Do you think you can make such a phage supply library?

Goodfellow: It is being done.

Mann: That is obviously of keen interest to us.

Hochstrasser: We have tried and we were partially successful.

Goodfellow: There are companies who are now running throughputs of 100 genes a month. You may extrapolate from that and say that it will therefore take 100 years to do the genome, but this is what we were saying about DNA sequencing not so long ago.

Venter: In a year or two we will have the complete genome. So you can start with the whole protein set.

In the proteomics field everybody seems to have a different definition of a unique protein. We have heard that there may be as many as half a million human proteins! Is there a consensus definition of what a protein is? Is a protein a new one because it has a phosphate on it, for instance?. I was invited to this lecture at the National Zoo after we sequenced the *Methanococcus* genome to talk about new species, and there was a fellow talking about all the different species of squirrels that he classified by slight changes in their tail. I don't see that as much different

than what you're doing with proteins. If you call a squirrel a new species because of a slight difference in the tail, is a protein a new one because it has a slight difference in its phosphorylation?

van Oostrum: It depends whether it has the same function or not. If it takes on a different function because it is phosphorylated, I would consider it a different protein.

Venter: That's the question: is it a structural definition or a functional definition?

Mann: One definition in the present context is that if a protein separates into a different location on a gel, then it is a different protein.

Goodfellow: I was interested in the attempts by Denis Hochstrasser to try to define the number of different variants per protein. You said that for *Escherichia coli* there was one spot of protein but it moved a bit, for yeast there were three spots, and in human there was six spots for each protein. Was that a theoretical calculation?

Hochstrasser: This is on the basis of comparing theoretical 2D gels and real 2D gels. We looked at the theoretical position of gene products and compared it to the real position of identified peptides. We matched the theoretical pictures to the real ones and found that in human sometimes you have three spots, sometimes 6 and sometimes 20 related to a single gene sequence.

Goodfellow: But you have only tested a few of all the possible spaces that these proteins could go to.

Hochstrasser: Yes.

Venter: Does proteomics have a role in the clinic?

Hochstrasser: Not yet. There are just a few areas where it is currently useful, such as in Creutzfeldt–Jacob disease (CJD) where you look at the presence of Tau chain. In CJD, there is no diagnostic test unless you do a brain biopsy. In the future, I believe that proteomics may play a crucial role in pathology. To illustrate why, I would say the following: when you do your rounds and you look at the patient and you have a biopsy report coming back from the pathology lab, you look first to see who signed the pathology report, because you are concerned about the expertise of the pathologist, which can be highly variable. So if we were to have a way of screening many molecules this would provide a much more accurate diagnosis and it would help the pathologist.

Venter: If that is the real world, how can we really go from genomics and polymorphic variation to understanding what the different drugs are going to do in the clinic. If the variation is so high that we can't measure it at the exo protein level, can we really extrapolate from the genome out?

Goodfellow: It's the same when you do any experiment. The first cut is easy, you do an experiment, and if you see something that changes then you focus on that and you don't worry about the other things which also change. I remember when p53 was discovered, it was discovered as a contaminating band on SV40 large T immunoprecipitations. This band had been seen for years, but was dismissed as a background contamination. That's how we do science: we take the easy stuff and then we go looking for the background bands.

There are many examples where we look at proteins for diagnostics. Once we know what protein we're looking for we set up a system which looks specifically for that protein or metabolite independently.

Venter: Can we do *ab initio* biology from the computer?

Goodfellow: I'm sure *you* can!

Mann: But it's difficult to get at unexpected mechanisms in that way.

Souza: One example that we have been using is that a small percentage of the genes that we get that are secreted, we get through our proteome project because the algorithms don't recognize the sequences as being secreted, compared to the database that we have generated. Yet using the proteome techniques you can collect glycosylated proteins (which by and large are those that are secreted) and find that you get different classes of proteins, you wouldn't get with the signal trap or the standard algorithm. Another example would be erythropoietin (EPO). You can make EPO as a *E. coli*-generated protein and it has normal activity in *in vitro* systems, until you put it *in vivo* and find it has no activity. This correlates with the level of sialic acid: the more sialic acid you can engineer into the molecule, the more *in vivo* potency it has.

Efcavitch: This goes back to comments made earlier: proteomics doesn't necessarily have to be used in the global sense, but rather for screening pathways, whether it's the metabolite modification that you're interested in monitoring, or whether it's a way of determining where that particular protein fits into a pathway. It is fun to look at all these global patterns, but perhaps the real power of proteomics is in

discrete pathway analysis or in looking for post-translational modifications of specific proteins.

Hochstrasser: When we will be able to identify all the spots and quantify and characterize them, it will make a big difference, because then we will be able to study epigenetic networks on a large scale. When it is just descriptive and you just look at an image it is not very useful. But things are moving rapidly in that direction.

Rubin: There is an immediate, very powerful use of proteomics that we may be losing sight of: that is for analysing complexes. For years, to do this we have been doing immunoprecipitation, running a Western blot, taking all the antibodies we have in the refrigerator and asking whether their target proteins are there or not. We had no way of assaying proteins for which we didn't have an antibody. There are a whole range of very interesting biological experiments that have already been published that need to be re-done looking at all the proteins in the complex. This is going to be an immediate, powerful use of these techniques.

Goodfellow: It's powerful when the *in vitro* technology that gave you the complex gives you function as well. For example, if you can perform DNA synthesis in the test-tube, and you can identify all the proteins that are in the synthesis complex, then you have a starting place to remove individual proteins to study their contribution to the complex.

Rubin: For example, Ras and Raf co-precipitate, but there are probably 30 proteins in that complex, of which we know about 10, because we have antibodies against them.

Goodfellow: I guess what I'm saying is that you would agree that once you have identified all the proteins, then you would either have to do genetics or you would have to have a functional biochemical test.

Rubin: I agree, but these methods are much more powerful than two-hybrid assays, which are limited to one-to-one interactions.

Mann: I'm glad that you say that, because that is exactly the way I see it. In a broader sense, you could say that with all the powerful sequencing-based approaches, we are forgetting about biochemistry. Now, with the mass spectrometry we are bringing back biochemistry. By taking out much of the tedious purification work of biochemistry with simple immuno-precipitations followed by mass spectrometric protein identification, we can get a short cut to function. In this way, we should be able to get at the function of many genes which you don't get from genome sequences alone.

Venter: To follow Gerry Rubin's point, how much can things be multiplexed? If you had the complete database of all the proteins, can you look without separation at 30 proteins at once?

Mann: A large-scale project we have been thinking about doing is to take the yeast genome, tag all the genes, and see what protein partners they precipitate. This would at least give us a list of the stable multi-protein complexes in yeast. We haven't done this yet because we believe we have even more interesting things to do, but it would give a lot of information for the money.

References

Efcavitch JW 2000 Electrophoresis-based fluorescent dideoxy-terminator sequencing. In: From genome to therapy: integrating new technologies with drug development. Wiley, Chichester (Novartis Found Symp 229), p 5–13

Fountoulakis M, Langen H, Evers S, Gray C, Takács B 1997 Two-dimensional map of *Haemophilus influenzae* following protein enrichment by heparin chromatography. Electrophoresis 18:1193–1202

Mann M 2000 Mass spectrometry resurrects protein-based approaches in functional genomics. In: From genome to therapy: integrating new technologies with drug development. Wiley, Chichester (Novartis Found Symp 229), p 27–32

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

Microbial genome sequencing: new insights into physiology and evolution

Claire M. Fraser

The Institute for Genomic Research (TIGR), 9712 Medical Center Drive, Rockville, MD 20850, USA

Microbes were the first organisms on earth and predated animals and plants by more than three billion years. They are the foundation of the biosphere—both from an evolutionary and an ecological perspective. The diversity of microbes, in terms of genetics, metabolism and physiology is far greater than that found in plants and animals. And yet, the diversity of the microbial world is largely unknown, with less than 0.5% of an estimated 2–3 billion microbial species identified (American Academy of Microbiology 1997). However, of those species that have been described, their biological diversity is spectacular, having adapted to grow under extremes of temperature, pH, salt concentration and oxygen levels.

Perhaps no other area of research has been so energized by the application of genomic technology than microbiology. Since TIGR published the first genome sequence for a free-living organism, *Haemophilus influenzae*, in 1995 (Fleischmann et al 1995) more than 20 other microbial genome sequencing projects have been completed. This progress has represented, on average, one completed genome sequence every two months and all indications point to this pace continuing to accelerate. Work is underway at TIGR and in other laboratories around the world on more than 60 microbial genome projects from a diverse group of pathogens, archaea and species of evolutionary importance (see *http://www.tigr.org* for a complete list). In the next 2–3 years, international efforts in microbial genome sequencing will generate more than 200 million base pairs (Mbp) of new DNA sequence containing ~200 000 predicted genes, at least 2–3 times the number of genes expected from the completion of the human genome project.

Work done to date has shown that there is tremendous variability in microbial genome size and GC content, ranging from a low of 29% for *Borrelia burgdorferi* (Fraser et al 1997) to a high of 68% for *Deinococcus radiodurans* (White et al 1999). The more than twofold difference in GC content has an effect on overall codon usage and amino acid composition among species. Genome organization is also variable with examples of single circular chromosomes, chromosomes plus one or a few plasmids or extrachromosomal elements, to the extreme seen with *B. burgdorferi*, a genome composed of a 910 kbp linear chromosome and 21 linear and circular extrachromosomal elements.

From a summary of results from the completed microbial genome sequences, representing more that 40 Mbp of DNA sequence and 40 000 predicted open reading frames (ORFs), it is immediately apparent that almost one-half of all ORFs identified to date are of unknown biological function (Table 1). Perhaps even more surprising is the fact that approximately one-quarter of the ORFs in each species studied to date are unique, having no significant sequence similarity to any other available protein sequence. Taken together, these data indicate that there is a substantial amount of microbial biology yet to be understood and suggest that the idea of a 'model organism' in the microbial world may not be valid, given the vast differences that we have observed, even between related species.

Other patterns are emerging with regard to proteins for which one can make putative assignments on the basis of sequence similarity searching. Within certain categories of genes, such as those involved in transcription and translation, for example, the total number of genes present in each genome is quite similar, even when genome size differs by fivefold or more. This observation suggests that a basic complement of proteins is absolutely required for these cellular processes. In contrast, the number of proteins in other functional categories, such as biosynthesis of amino acids, energy metabolism, transporters and regulatory functions, for example, is more variable and tends to increase as genome size increases. Thus, as genome size increases so too does biochemical complexity for a given organism.

A significant proportion of larger microbial genomes represents paralogous genes, that is, genes related by duplication rather than by vertical inheritance. As shown in Table 2, the number of genes that are

TABLE 1 Summary of features from completed microbial genomes

Organism	*Genome size (Mbp)*	*Number of ORFs*	*Unknown function*	*Unique ORFs*
A. fulgidus	2.18	2437	1315 (54%)	641 (26%)
M. thermotautotrophicum	1.75	1855	1010 (54%)	496 (27%)
M. jannaschii	1.66	1749	1076 (62%)	525 (30%)
P. horikoshii	1.74	2061	859 (42%)	453 (22%)
A. aeolicus	1.50	1521	663 (44%)	407 (27%)
B. subtilis	4.20	4100	1722 (42%)	1053 (26%)
B. burgdorferi	1.44	1751	1132 (65%)	682 (39%)
C. trachomatis	1.04	894	290 (32%)	255 (29%)
D. radiodurans	3.28	3192	1715 (54%)	1001 (31%)
E. coli	4.60	4288	1632 (38%)	1114 (26%)
H. influenzae	1.83	1692	592 (35%)	237 (14%)
H. pylori	1.66	1657	744 (45%)	539 (33%)
M. tuberculosis	4.41	3924	1521 (39%)	606 (15%)
M. genitalium	0.58	470	173 (37%)	7 (2%)
M. pneumoniae	0.81	677	248 (37%)	67 (10%)
Synechocystis sp.	3.57	3168	2384 (75%)	1426 (45%)
T. martima	1.86	1877	863 (46%)	373 (26%)
T. pallidum	1.14	1040	461 (44%)	280 (27%)
	39.25	38 353	17 782 (46%)	9910 (26%)

TABLE 2 Summary of paralogous genes

Organism	*Genome size (Mbp)*	*Number of ORFs*	*Paralogous ORFs*[a]
T. pallidum	1.14	1040	129 (12%)
B. burgdorferi	1.44	1751	707 (40%)
H. pylori	1.66	1657	266 (16%)
A. fulgidus	2.18	2437	719 (30%)
B. subtilis	4.20	4100	1947 (47%)
M. tuberculosis	4.41	3924	2000 (51%)
E. coli	4.60	4288	2272 (53%)

[a]ORFs that share at least 30% sequence identity over more than 60% of their lengths.

contained in paralogous gene families increases as genome size increases. The one exception to this rule is seen with *B. burgdoferi* (Fraser et al 1997), but this organism is unusual in that it contains a large number of plasmid-encoded lipoprotein paralogues. The largest classes of paralogues in essentially all genomes studied to date are the ATP-binding proteins associated with ATP binding cassette transporters.

The availability of more than 20 completed microbial genome sequences have provided new insights on microbial evolution and diversity. The molecular picture of evolution for the past 20 years has been dominated by the small subunit ribosomal RNA phylogenetic tree of Carl Woese that proposes three non-overlapping groups of living organisms, the bacteria, the archaea and the eukaryotes (Woese & Fox 1977). Although the archaea possess bacterial cell structures, it has been suggested that they are no more closely related to bacteria than to eukaryotes. This three domain proposal also posits that the archaea and the eukaryotes shared a common ancestor exclusive of bacteria, or in other words, the common ancestor of eukaryotes descended directly from within the archaeal lineage.

As a result of the completion of genome sequences from representatives of all three domains of life, it is now possible to examine evolutionary relationships among living organisms in a more comprehensive way. However, this task has turned out to be anything but straightforward. Incongruities can be seen everywhere in the phylogenetic tree from its root to the major branchings when single protein phylogenies are examined. It has become clear that gene evolution does not equal species evolution. This, in large part, is a result of extensive lateral gene transfer, not only between bacteria but also between bacteria and archaea (Nelson et al 1999).

Beyond trying to decipher molecular evolution, another formidable challenge in microbial genomics will be how to make use of the new sequence information on a large scale to better understand biology. By using approaches that include gene chips, microarrays and proteome analysis it should be possible to move from a static picture of a genome, as captured in a set of DNA and protein sequences, to an identification of gene networks and a better understanding of the dynamic nature of the regulation of gene expression in the microbial cell.

References

American Academy of Microbiology 1997 The microbial world: foundation of the biosphere? American Academy of Microbiology, Washington, DC

Fleischmann RD, Adams MD, White O et al 1995 Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Fraser CM, Casjens S, Huang WM et al 1997 Genomic sequence of a Lyme disease spirochete, *Borrelia burgdorferi*. Nature 390:580–586

Nelson KE, Clayton RA, Gill SR et al 1999 Genome sequencing of *Thermotoga maritima*: evidence for lateral gene transfer between archaea and bacteria. Nature 399:323–329

White O, Eisen JA, Heidelberg JF et al 1999 Genome sequence science of the radioresistant bacterium, *Deinococcus radiourans* R1. Science 286:1571–1577

Woese CR, Fox GE 1977 Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 74:5088–5090

DISCUSSION

Hoffman: One of the issues that comes up in terms of trying to organize support for comparative genomics is cost. In malaria it has been suggested that because of cost we should take a genes-first approach to sequencing the additional *Plasmodium* sp. genomes after we complete the sequence of *P. falciparum* and not sequence the intergenic regions at all. On the basis of what you have learned from your comparison of genomes, what do you think we will lose by only focusing on genes?

Fraser: First, perhaps a point of clarification: one of the real advantages of microbial genome sequencing is that essentially all microbial genomes consist of wall-to-wall genes. Greater than 90% of the sequence in prokaryotic species represents coding sequence. Therefore, in bacteria I don't think there's any advantage to be gained by going after just the genes and leaving out the intergenic regions. With some of the more complex species, I think there's a great deal that we can lose by ignoring the intergenic regions. Where we are today with the efficiency and cost-effectiveness of shotgun sequencing, there is no real benefit to be gained from going after the genes only, and a lot of valuable information is lost.

Venter: Picking up on a point Dalia Cohen raised earlier, if you don't have all the genes in a family, you don't know whether you are pursuing the right target or not. Having 80% of the genes sounds great, but if the one target you're looking for is in the 20% that's missing, it doesn't help you very much.

Rubin: I have a question about the 50% of genes in bacteria that don't have an obvious function. In your analysis, do these genes tend to be

unique to one organism, or do you find them in many species? Among those that exist in multiple species, do they tend to be limited to a particular type of bacteria?

Fraser: There are examples of genes that are conserved fairly widely, that is, those that are found in a large number of species, and others with a more limited species distribution.

Venter: There was a very small subset of genes that occurred in all the species we sequenced: 200 or so were found in every species.

Fraser: I don't think the number is that high, particularly if you're looking at both bacterial and archaeal species.

Venter: The question is, are the species-specific ones just an artefact of the small sampling we have currently? They may not be species-specific as we get more genomes.

Rubin: A key question is how is one going to determine the function of all these genes and how should one prioritize one's efforts? I think it is more important to go after the ones that are widely distributed, because they may have general functions that we don't know anything about. On the other hand, if you are looking for a drug target, it is probably better to pick those that are species specific.

Fraser: I agree; it depends on what you're interested in. In terms of therapy or potential drug targets, the criteria you would select would be very different from those you would choose if you were trying to understand biology that was shared by a large number of species. This is where the power of comparative genomics really comes in, in enabling us to begin to categorize genes according to these criteria.

Venter: It is best to go after the low-hanging fruit first. If you are looking at drug targets, it's much better to go after things you can recognize at the present time. The trouble is if you are writing an NIH or MRC grant, if you don't have a hypothesis about what these are, you won't get funding in the first place. It is going to be a real problem, because we can't do descriptive biology at the stage where genomics demands that.

Rubin: The big problem is the cost. Although the cost of sequencing is coming down, the cost of determining functions of genes for which you just have a sequence and know nothing else of, has not reduced dramatically over the last 20 years. Even for determining the function of an unknown open reading frame in *Escherichia coli* you are looking in terms of person years.

Venter: Diversa is looking for new enzymes, as an example, and they have set up a number of high throughput screening assays for new enzymatic functions. So there are some approaches that could be used, but nothing systematic that I'm aware of in terms of specific gene-by-gene function.

Goodfellow: With the sequences of the genomes available, do we still have a concept of species?

Fraser: A good question. This has become more difficult to sort out with the wealth of genomic data, rather than becoming more obvious. It appears that lateral gene transfer plays a tremendous role in generating diversity in microbial species. Our recent paper in *Nature* on the *Thermotoga* genome showed that fully one-quarter of the genes in this organism are most similar by far to archaeal genes (Nelson et al 1999). These are not genes that are scattered along the chromosome—they are large pieces of DNA, in some cases flanked by repeat sequences, that look as if they may have been acquired via gene transfer of unknown mechanism. It really makes us stop and think about what a species is, and what it isn't. If evolution in the microbial world is more dominated by lateral gene transfer than by vertical descent of various genes, some of the differences that we are seeing as we begin to look at species that are closely related are making it more difficult to define exactly what a species is.

Venter: Species definitions certainly become much more complicated. We think that the Woese tree of life is not the correct picture. It is going to turn out to be much more of a neural network-type mesh which makes species definitions remarkably difficult, except for complex organisms—after all, we differ from cows only slightly.

Goodfellow: It astonished me to hear that only one base pair in every 3200 is different between two strains of *Mycobacterium* which have been separated for a huge number of generations. Isn't that remarkable? Essentially the same experiments have been done in humans, and the answer comes out at one base pair every 1000.

Fraser: Actually, in the tuberculosis community there was a great deal of surprise at how different these two strains turned out to be, on the basis of some studies from Jim Musser's group in Texas, which had looked at polymorphisms in a limited set of genes in a number of isolates of *Mycobacterium tuberculosis* (Sreevatsan et al 1997). The conclusion was

that there were few differences among strains of *M. tuberculosis*—far fewer than we observe from whole genome analysis.

Venter: M. tuberculosis seems to be one of the most conserved genomes, but it is not totally clear why.

Goodfellow: Is it a constraint of the GC content?

Fraser: That could be part of it.

Lipshutz: There is some work that Tom Gingeras did looking at *M. tuberculosis* and also a number of other *Mycobacterium* species (Gingeras et al 1998). This is just looking at about three or four different genes. While there is a fair bit of variation between variants in other species such as *M. avium*, *M. tuberculosis* was surprisingly conserved—much more so than any other species of *Mycobacterium*.

Fraser: Yes, in terms of what has been done previously, the results were entirely unexpected.

Goodfellow: This low level of variation is inconceivable to me.

Rubin: I know in *E. coli* there is supposed to be a very large difference between lab strains and clinical isolates. What is the general situation with other bacteria, so we can put this in context?

Fraser: In *E. coli* and *Helicobacter pylori*, where the comparisons are possible at the whole genome level, the differences between strains are much greater. The overall chromosome organization in *Mycobacterium* looks very much the same between strains, but this isn't the case in *E. coli* and *H. pylori*. This isn't seen to the same extent even with *Chlamydia*.

Venter: In this context, the best two genomes to look at for the purposes of comparison are probably those of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. Even though the *M. genitalium* genome is completely contained in the *M. pneumoniae* genome in terms of gene content, they are only about 30–50% identical at sequence level. *M. pneumoniae* has 200 extra genes, but the gene sequence variation is tremendous. One of the things with the mycobacteria is that perhaps the situation is a little confusing: people thought the Oshkosh strain was a new emerging *Mycobacterium*. But we have begun to think that this may be an ancient *Mycobacterium* that has actually come back, and so the reality is that we may not be looking at strains that are all that far apart in time.

Goodfellow: But you can do the calculations as well as I can in terms of the number of generations that human beings have been on this planet,

compared with the number of generations that *Mycobacterium* goes through in one infection. I just don't understand this lack of variation—I think it's remarkable.

References

Gingeras TR, Ghandour G, Wang E et al 1998 Simultaneous genotyping and species identifcation using hybridization pattern recognition analysis of generic *Mycobacterium* DNA arrays. Genome Res 8:435–448

Nelson KE, Clayton RA, Gill SR et al 1999 Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature 399:323–329

Sreevatsan S, Pan X, Stockbauer KE et al 1997 Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. Proc Natl Acad Sci USA 94:9869–9874

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

Pharmacogenetics and pharmacogenomics in the discovery and development of medicines

Allen D. Roses

GlaxoWellcome Research and Development, 5 Moore Drive, Research Triangle Park, NC 27709, USA

The terms 'pharmacogenomics' and 'pharmacogenetics' are often interchanged and used without clear definition. For the purpose of this Novartis Foundation Symposium, I will use working definitions. Pharmacogenetics refers to *people* including gene identification and 'right medicine for right patient'. Pharmacogenomics refers to the application of *tools* including, but not limited to, the functional genomics toolbox of differential gene expression, proteomics, yeast two-hybrid analyses, tissue immuno- and histopathology, etc.

There are two applications of pharmacogenetics that may use similar techniques but are quite distinct: susceptibility gene identification and 'right medicine for right patient'.

Susceptibility gene identification

For monogenic diseases, current linkage methods are now extremely efficient in identifying mutant genes, depending mostly on the total amount of family structures and DNA samples available. For susceptibility genes, identification of confirmed polymorphisms associated with the disease have been much more challenging. In general, a comparatively large linkage area with indistinct boundaries has been the best scientists can provide. Within these large linkage areas there may be hundreds of genes that are usually examined one at a time for candidate gene association. While there are many candidates, each with a proposed relationship to the disease, very few widely confirmed

susceptibility gene identifications exist. The apolipoprotein E locus (*APOE*) association with common, late-onset Alzheimer's disease (AD) was the first polymorphic susceptibility locus identified by linkage for a major disease. The association of the *APOE4* allele with earlier age of onset distributions, and thus increased risk, was confirmed in over 150 populations with no non-confirmations in any group of more than 30 patients and controls. The association of the *APOE2/3* genotype with a later age of onset and decreased risk is also widely confirmed. Thus common *APOE* genotypes carried by people can be interpreted in multiple populations in epidemiological models.

To test whether or not high-density single nucleotide polymorphism (SNP) mapping could detect a susceptibility locus within a large region, GlaxoWellcome scientists constructed a SNP map of 2 megabases (mb) on either side of *APOE* (Lai et al 1998). We asked the question whether a SNP map analysis could detect the location of the *APOE* locus for AD, if we did not know it was there. The locus was narrowed to less than 100 kilobases (kb), which included the *APOE* locus, in a very short time frame. This process has since been employed within GlaxoWellcome for other disease susceptibility gene searches through large linkage regions, including psoriasis, diabetes mellitus, migraine, chromosome 12-linked AD and others. These experiments will define the practical density of SNP maps useful for narrowing the large linkage areas to 50–200 kb, containing far fewer candidate genes that could then be tested for disease association (Martin et al 2000).

The construction of a whole genome high-density SNP map clearly focuses the next stage of susceptibility disease gene research on the availability of well-constructed, accurately phenotyped patient populations. In anticipation of The SNP Consortium (TSC) map, GlaxoWellcome is generating useful patient collections from multiple diseases with large unmet medical need.

'Right medicine for right patient'

Can we use genetic profiling to recognize patients who will respond positively to a particular medicine? Can we use profiling to identify those patients who will have an adverse event by taking a particular medicine? Can genetic profiling be performed at reasonable cost using a

standardized genetic map? These were some of the questions that led to the formation of TSC.

Assuming that a whole genome SNP map with a density of 15 kb average were to be used, this would be approximately 200 000 SNPs. Each SNP genotype would require at least two reactions, one for each allele, or 400 000 genotypes per person. In a phase II trial with 500 people of whom 100 were drug responders, 200 million genotypes would be required. This one experiment, if based on a cost of US$0.01 each, would cost $2 million. Clearly for these experiments to be affordable for development of early phase drugs, the cost and speed of genotyping will need to be significantly different than today's costs and methodologies allow. Our current data and future experiments will determine the practical SNP density that will be needed to profile patients. Although it is estimated that there may be several million potential SNPs in the human genome, the practical significance of a commercial experiment must be a consideration. The current goal at GlaxoWellcome is to be able to measure 200 000 SNPs in 500 people over a two-week period at a reasonable cost, since we perform in excess of 25 such clinical trials annually. We have developed a bead-based system that has been beta-tested in parallel with standard methods of SNP analyses.

For a SNP mapping system to be useful across the industry, particularly with regulatory authorities, it must be standardized, readily available and amenable to GLP procedures. It is expected that profiles of SNP linkage disequilibrium maps could be abstracted down to several hundred to a few thousand SNPs and be analysed using conventional chip methodologies. If a SNP profile were to be useful when linked to a medicine prescription, then hundreds of thousands of conventional chips would need to be distributed to diagnostic laboratories. **It is important to note a critical ethical point: the abstracted SNP profile would give no information concerning any genetic characteristic other than the medicine response, and thus no collateral information to family members concerning any genetic disease would be available** (*http://207.78.88.3/fda/transcripts/tran_roses.htm*).

The time-frame for the SNP map is two years, with concurrent development of analytical methods and bioinformatic (data-mining) read-out methodologies. Application to medicines that are already

registered and in the market place, but have significant adverse characteristics that limit their commercial value, will no doubt be the first area studied over the next five years. These studies will also provide the proof of principle for parameters for registration of new medicines during the next five years.

References

Lai E, Riley J, Purvis I, Roses A 1998 A 4-Mb high-density single nucleotide polymorphism-based map around human APOE. Genomics 54:31–38

Martin ER, Gilbert JR, Lai E et al 2000 Analysis of association at SNPs in the APOE region. Genomics 63:7–12

DISCUSSION

Hoffman: Malaria is reputed to have driven the human genome more than any other infection, and there are a number of haematological disorders associated with susceptibility to malaria. In designing field trials and vaccines, we would very much like to know which individuals are more susceptible to malaria and which are protected, to limit the size of the studies. The sickle cell trait confers approximately 90% protection against death from malaria and has a penetrance in the population of about 10%. In many parts of sub-Saharan Africa, the infant mortality rate is about 10%, and up to half these deaths may be due to malaria. Thus 5% of infants may die of malaria, and amongst those that die, less than 1% will have the sickle cell trait. Amongst those that live, at least 10% will have sickle cell trait. In the same way as you did for AD, could you set up a SNP analysis type of approach that will actually find the sickle cell trait, validate the methodology, and then go from there to perhaps look at the flip side for the people that won't die from malaria using a similar type of analysis.

Roses: I haven't thought a lot about malaria, but I can tell you about tuberculosis and AIDS. There is a group of prostitutes in Africa who clearly get exposed to AIDS but have never come down with it, and we would like to know why. By collecting DNA from them and comparing them with the rest of the population, one might the have a way of profiling them for multiple SNP variants that are in linkage

disequilibrium. Similarly, in the Gambia and South Africa, there are clever ways of studying tuberculosis. In the Gambia there is a high prevalence of twinning, and there are some interesting studies which have looked at twins in which one gets tuberculosis and the other doesn't. In the general population as well, there are people who get exposed because they're in families with tuberculosis and they don't get it. We can pick out those people by SNP profile analysis. I suspect that you would be able to do this with malaria and the sickle cell trait also.

Venter: The statistics on sickle cell are probably far greater than in the study on AD, although he started with a narrow region when he did a SNP analysis: it was not a whole genome analysis because the linkage to a specific region was already done.

Roses: I suspect that if we did that in that experiment, you would light up the area around several regions of the genome. Then you would have to figure out what the polymorphism was that led to the sickle cell.

Efcavitch: I'm curious about your use of 200 000 SNPs in a phase II clinical trial. This seems awfully late in the development process for so many SNPs. 200 000 sounds like an association study as opposed to pharmacogenetics.

Roses: It is association, if you will, but it is not an association of a haplotype, in which you can take one polymorphism from a location on one chromosome and a second polymorphism from another chromosome, and you have to multiply it among hundreds of people involved. The thing about the SNP map is that it is constrained in an order. As you scan through the population you're asking whether these ordered SNP profiles are really close enough together to be able to detect linkage disequilibrium between multiple SNPs. The simulation given in Kruglyak (1999) says it has to be 3 kb apart in order to detect linkage equilibriums. This estimate doesn't seem to match the data in the disease loci that we've looked at.

Venter: Identifying 200 000 SNPs doesn't suggest a knowledge-based approach. If you knew the variation causing the problem for those individuals, you might measure just five.

Roses: You would have me measure three million!

Venter: Actually, I wouldn't. I would use the three million to get down to the few that actually make sense, instead of trying to do a blind thing across 200 000.

Efcavitch: That's really the key question. Again, I'm only asking about this in the context of phase II studies, as opposed to further upstream in the development process from phase II studies.

Roses: The first use of this would be to take drugs for which there is something that keeps the drug from being useful—such as a side effect. For example, we have a drug on the market for which the problem is that a certain percentage of the people get unpleasant skin rashes, so it is not widely used. During clinical trials dose escalations were employed so we could avoid the skin rashes, and as soon as you put in dose escalations, family doctors are not going to use the drugs because they only have a few minutes to see each patient. If we could pick out the people who are not going to develop the skin rash by a SNP profile (abbreviated to only include informative SNPs), then testing with a couple of hundred SNPs would allow selection of people who would respond adversely. Without even figuring out the mechanism behind the skin rash or the genes involved, the abstracted or abbreviated SNP profile could allow you to take a drug that otherwise would require dose escalations, and avoid complications.

Venter: If your linkage disequilibrium works with that number of SNPs, you might get down to the actual cause of the side effect.

Roses: We'd obviously follow that up.

Efcavitch: You are talking about 20 million SNPs a day, which is a daunting technological challenge. It is still an open question of economics for the drug recovery model versus targeted pharmacogenetics where one has known sites of action or candidate genes, where one is talking about a much smaller number of SNPs.

Roses: Let me put this in perspective. That's a US$400 million drug per year. The estimate is that if you could do this it would become a billion per year.

Venter: So it makes economic sense to pay for SNP analysis.

Mann: How do you score the SNPs now, and how will you score them in the future? And how accurately do you have to score each one?

Roses: The power changes every three months. We have developed some microsphere-based approaches, using principles of combinatorial chemistry. We have used Luminex® beads that are different colours and have developed a rapid, relatively inexpensive method of scoring and running SNP analysis. It looks like it's going to work. It would be less

expensive if PCR could be avoided. Technology companies are beginning to come out with methods that might be applicable, such as third wave methods, without using PCR. We have to wait two years for the full SNP map. There are a lot of smart people who are no longer spending their time trying to generate SNPs because it is going to be done anyway, but instead are trying to figure out how to read SNPs. With regard to the bioinformatics of reading the SNPs (we are currently publishing a paper with SAS, Duke and NC State on this), this will be commercially available. The profiling and bioinformatics is therefore not a big problem. However, collection of the patient population is the major problem. Once the technology is there, it is the accuracy of phenotypic characterization that will determine the usefulness of this approach. In clinical trials, doctors get paid to examine these patients, and they have to fill out very rigid data forms, so these people are about as well phenotyped as any group we are going to get. As a proof of principle, we are looking at 16 different molecules in thousands of clinical trial patients. It is estimated that we and the company could save hundreds of millions of dollars a year if we could significantly focus and cut the expense of phase III clinical trials.

Hochstrasser: A quick question about APOE. If I understood correctly, the mouse doesn't have APOE, but you added the human *APOE* gene and found no difference in the phenotype.

Roses: There are two different mouse types. One set are *ApoE* knock-ins, which lack intraneuronal expression, similar to wild-type mice. The other mice have the *ApoE* knockout background and the human *APOE* genetic fragments. These mice show intraneuronal APOE, as in humans. These are used in two distinct parallel screens.

I don't understand why people actually study a knockout for *ApoE*. There are only three humans reported in the world literature who didn't have APOE. All humans have APOE—they have different types, and depending on the types you get AD earlier than normal. The nuance of studying something that has APOE3 versus APOE4 is more important than studying something that lacks APOE.

Venter: Can you use the SNP data collected from patients for other studies?

Roses: All these studies are very well co-ordinated through the pharmaceutical companies and medical people who are on top of each of

the studies. What we have done is integrated into our international development groups, so that every medical development group has a genetics component.

Venter: But can you use the SNP profiles for other diseases, or other studies? Do you have their consent to do this?

Roses: Every one of them are consented for research use for commercial purposes.

Reference

Kruglyak L 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22:139–144

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

Mutagenesis and genomics in the mouse: towards systematic studies of mammalian gene function

Steve D. M. Brown

MRC Mammalian Genetics Unit and UK Mouse Genome Centre, Harwell, Oxfordshire OX11 0RD, UK

Mouse mutagenesis for systematic studies of gene function

Systematic approaches to mouse mutagenesis will be vital for future studies of mammalian gene function. However, mouse mutants are available for only a small percentage of the total number of mammalian genes—there is a 'phenotype gap' (Brown & Peters 1996). We need to generate more mouse mutations in order to increase both the breadth and depth of the mouse mutant resource, recovering mutations at new loci as well as identifying new alleles at known mutant loci. This process can be phenotype- or genotype-driven, and both approaches have a role to play in delivering new drug targets. Genotype-driven mutagenesis often underpins target validation approaches. The advantage of genotype-driven approaches (e.g. gene trap embryonic stem cell libraries) is the ease of identification of the mutated locus; the disadvantage is that prior assumptions often have to be made concerning the likely function and phenotype of the mutated locus. In contrast, the phenotype-driven approach makes no assumptions about the underlying genes involved and emphasizes the recovery of novel phenotypes. One phenotype-driven approach that is playing an important role in expanding the mouse mutant resource employs the mutagen *N*-ethyl-*N*-nitrosourea (ENU) (Brown & Nolan 1998).

ENU mutagenesis: phenotype-driven mouse mutagenesis screens

At Harwell, and in collaboration with colleagues at SmithKline Beecham, Imperial College and the Royal London Hospital, we have begun a major

ENU mutagenesis programme incorporating a large genome-wide screen for dominant mutations. Over 18 000 mice have been produced to date and the majority screened employing a systematic and semi-quantitative screening protocol—SHIRPA (Rogers et al 1997). SHIRPA is a hierarchical screening protocol employing a rapid and efficient primary screen for deficits in muscle and lower motor neuron function, spinocerebellar function, sensory function, neuropsychiatric function and autonomic function. Moreover, in the primary screen blood is collected from all mice and subjected to a comprehensive clinical chemistry analysis. Subsequently, secondary and tertiary screens of increasing complexity can be employed on animals demonstrating deficits in the primary screen.

Frozen sperm is archived from all the male mice passing through the screen. In addition, tail tips are stored for DNA.

Progeny testing of mice carrying abnormal phenotypes indicates that around 1–1.5% of mice from the screen carry a new heritable dominant phenotype. Nearly 100 mutants have been confirmed as heritable and added to the mouse mutant catalogue. (For further information on the project and details of data derived from the screening see: *http://www.mgc.har.mrc.ac.uk/mutabase/*).

Creating the mouse mutant map

Overall, the ENU mutagenesis programme will provide an extensive new resource of mutant and phenotype data to the mouse and human genetics communities at large. The challenge now is to employ the expanding mouse mutant resource to improve the mutant map of the mouse—and for this it is necessary to devise rapid strategies to genetically map new mutants. We are currently using frozen sperm and IVF for the rapid generation of small backcrosses in order to map many of the newly catalogued mutations to the mouse genome (Thornton et al 1999). Nevertheless, despite the availability of semiautomated genotyping approaches for genetic mapping, this particular step remains a bottleneck for the rapid development of the mutant map. The development of mouse genotyping chips will significantly enhance the rate of progress of mutant mapping.

Harnessing the mutant map to ongoing genomics programmes

As the mouse mutant map develops there needs to be a commensurate improvement in the mouse gene map that will be delivered via programmes such as expressed sequence tag (EST) mapping and comparative sequencing. An international programme is underway to generate a dense EST map of the mouse using a mouse Radiation Hybrid mapping panel (McCarthy et al 1997). A large number of unique embryonic and tissue cDNA libraries have been developed in the mouse and have been used to generate large numbers of ESTs not so far identified in human. Assignment of these ESTs to the mouse map will significantly enhance the mammalian gene map and in so doing improve the identification of candidate genes for loci on the mouse mutant map.

A draft human genome sequence is expected by year 2000. Plans to provide a draft sequence of the mouse genome have been initiated. In addition, the provision of finished sequence from several defined regions of the mouse genome is already underway. Comparison of human and mouse sequence in any region is expected to improve the identification and annotation of gene sequences and provide an important adjunct to gene prediction software. Indeed, in at least a few cases to date, the provision of mouse and human comparative sequence has underpinned the identification of novel genes and their mutation scanning. One recent example is the identification of the mouse X-linked *Bare patches* (*Bpa*) and *Striated* (*Str*) mutations (Liu et al 1999), both dominant male lethals having pleiotropic effects on skin morphology and skeletal development. Comparative sequencing of the region in which the *Bpa* and *Str* mutations were known to lie aided the characterization and annotation of a novel 3β-hydroxysteroid dehydrogenase gene, *Nsdhl*. Subsequent mutation analysis demonstrated that *Bpa* and *Str* were allelic mutations within this gene. *Nsdhl* appears to play an important role in cholesterol biosynthesis and the association of mutant phenotypes with lesions in this gene expands the spectrum of phenotypes associated with abnormalities of cholesterol metabolism.

Conclusion

The development of an improved mutant map of the mouse will be an important asset in exploiting the growing gene map of the mouse and

assisting with the identification of genes underlying novel mutations, with consequent benefits for the analysis of gene function and the identification of novel pathways. The delivery of a new mouse mutant catalogue along with the resources for rapid gene identification will bring noticeable benefits for the identification and characterization of novel drug targets.

References

Brown SDM, Nolan PM 1998 Mouse mutagenesis—systematic studies of mammalian gene function. Hum Mol Genet 7:1627–1633

Brown SDM, Peters J 1996 Combining mutagenesis and genomics in the mouse—closing the phenotype gap. Trends Genet 12:433–435

Liu XY, Dangel AW, Kelley RI et al 1999 The gene mutated in *Bare patches* and *Striated* mice encodes a novel 3β-hydroxysteroid dehydrogenase. Nat Genet 22:182–187

McCarthy LC, Terrett J, Davis ME et al 1997 A first-generation whole genome-radiation hybrid map spanning the mouse genome. Genome Res 7:1153–1161

Rogers DC, Fisher EMC, Brown SDM, Peters J, Hunter AJ, Martin JE 1997 Behavioral and functional analysis of mouse phenotype: SHIRPA, a proposed protocol for comprehensive phenotype assessment. Mamm Genome 8:711–713

Thornton C, Brown SDM, Glenister P 1999 Large numbers of mice established by *in vitro* fertilization with cryopreserved spermatozoa: implications and applications for genetic resource banks, mutagenesis screens and mouse backcrosses. Mamm Genome 10:987–992

DISCUSSION

Lipshutz: How are you doing the low resolution mapping?

Brown: We take our founder mice (or progeny from the founder mice), we archive the sperm and can do IVF in a dish, quickly generating hundreds of progeny. Initially, for mapping the mutants we are generating in the order of 100–150 backcrossed progeny. Then we can quickly populate the shelves in a matter of weeks with those progeny, take tails and screen. With SmithKline Beecham we've developed a panel of a couple of hundred microsatellite markers around the genome for the two strains. Following pooling we are doing PCR and standard automated fluorescent genotyping. This is working very effectively. However, I should say that in terms of capacity, one of the important things to look to for the future is to move to a faster genotyping route in the mouse through single nucleotide polymorphisms, for instance. This is something that we are looking at. Currently there are only about a thousand mouse mutants in the database. This will likely increase by three- or fourfold over the next three years, and the value of that mutant

resource is going to lie in getting those onto the genetic map rapidly and at high resolution. Using IVF has sorted out the bulk bottleneck of doing the crosses, but we could do with better genotyping approaches.

Venter: What's the availability of those data sets?

Brown: All the data are public and the mice are available through an MTA (Material Transfer Agreement). The mice are freely available to academics, and indeed many mice are going out to academic institutes around the world. In addition, we have a number of collaborations with other centres: people are coming in with particular screens to try to tease out other phenotypes from the mice that are being produced. This is quite important for us. We have tried to 'hotel' the facility as much as possible, so that people can come in with novel imaginative screens to look for particular phenotypes.

Venter: Do you maintain all the mice, or do you just maintain the frozen sperm?

Brown: At the moment we're maintaining all the males plus abnormal phenotypes that we have detected as they arise in the female. We are actually maintaining all the male progeny that are coming through irrespective of whether we see a phenotype or not. Once we get the licences in place, we are about to implement ovary freezing for all the females. The idea is that ultimately we will retain all the mice. This is important in terms of thinking about data mining: going back to the data, teasing out new phenotypes and then being able to recover the mouse and examine it in more detail.

Venter: So if the mouse genome is available in 12–18 months, how will you relate all your data back to the genome?

Brown: This is where the mapping of the mutants is very important. We want to get fast, relatively high resolution mapping in position, and to be able to look for candidate genes.

Venter: Are you being funded to do that?

Brown: Yes, in particular areas, where we have particular programmes of interest. For instance, my own special interest is in the genetics of deafness, and we have funding to go down this route.

Venter: But you are not funded systematically?

Brown: No. To some extent, I see this as a long-term resource for the mouse community at large. We aim to have a large number of mutants available for years to come, where people would go in and pick out a new

mutation which becomes interesting for whatever reason. Our perspective at Harwell is that mutants that have lain dormant in our embryo bank for 20 years or more suddenly become interesting for a particular reason and there are suddenly many requests for them. I'm sure the same will be true for this new mouse mutant resource.

Goodfellow: It might be possible to perform saturation mutagenesis and freeze sperm from mutation-carrying animals. Stored DNA samples could then be used to search for specific mutations. I'm sure there will come a time over the next decade where you will have 'dial a mutant'—if someone wants a mutation in any particular gene, they can go and screen stored DNA and get the answer back the next day. If you do the calculations, you only need something like a million mutated animals to have a mutation in every codon, so it's not beyond the technology that we have today.

Brown: I agree that this is feasible. I didn't mention the genotype-driven approach to go out and look for mutations in specific genes. However, I do like the phenotype-driven approaches as well, because you are looking for phenotypes of particular pathways, where you're actually making no a priori assumptions about which genes are important in that pathway. Whereas if one goes and looks for a particular gene, there are two problems: first, you're making an assumption about why this is an interesting gene, and second you are also making assumptions about the phenotype that you should look for in that mutant mouse. This is not a trivial problem. We all know the problems of looking for phenotypes in knockouts, partly because people make assumptions about what is the role of a particular gene. The two approaches are definitely complementary, and both need to be driven forward.

Roses: How would you apply this to a complex disease where you will not necessarily see a mutant phenotype?

Brown: Effectively, we're generating monogenic models here.

Venter: How do you know that?

Brown: OK, I'll backtrack on that statement to some extent. Of course there are many hits around the genome. There is a specific locus mutation rate of one in a thousand. There must have been around 50–100 hits, many of which must be silent. In all the inheritance testing that we have done we haven't got an example yet of different aspects of a phenotype segregating out. To some extent, this just says how much of the genome is relatively

silent in our ability to pick up phenotypes in the mouse. Of course, that does not mean there are not phenotypes there. By and large the major phenotypes we find look like monogenic traits. I think the future of mouse genetics in the next 10–20 years is things like modifier screens, making compound mutants, and being able to effectively generate what we might call a polygenic situation.

Goodfellow: The chance of modelling any particular polygenic disease in any particular strain would be very small. You are starting with inbred strains of mice which are fixed in particular allele sets, and everybody knows who works with inbred strains of mice, that they're very different from each other.

Venter: What is the polymorphic rate in Balb/c mice?

Brown: I doubt that it is zero, but it would be very low.

In a sense I agree with Peter Goodfellow's point: if you generate anything on any genetic background you can say it's polygenic, but the same mutant generated on a different inbred strain might not be exactly the same phenotype.

Goodfellow: And that's what we see. In fact, a big problem with studying complex genetic traits in mice is that often when you do a cross between different inbred strains, you get a different answer.

Rubin: How much of a limitation is it that you're limited to looking at dominant phenotypes?

Brown: Obviously, it is a limitation in some sense. Genome-wide recessive screens are being planned, but this can't be done on the same scale. There are three generation screens. The approach that will be more significant in the mouse in terms of getting recessive alleles is to use deletion or inversion screens to particular mutants.

Rubin: Do you have any idea as to how many of the mutations you are getting are due to haploinsufficiency?

Brown: We have no idea yet.

Venter: So if we're going to sequence a strain of mice, it seems that Balb/c or any other inbred strain would represent a poor choice. We should instead pick a street mouse, to have the polymorphic variation to work back to linkage of the traits.

Brown: Any reference strain would do. Of all the different inbred strains that people will be using, including wild-mouse variants as well, people will be re-sequencing to look at the polymorphisms and the variation.

Kopczynski: It seems to me that you are screening enough that with the recessive frequency of mutations you must be coming up with the same hits several times. Can you look at the phenotypes and make predictions? For example if you find the same cranofacial mutation three times, can you predict that it is a loss-of-function?

Brown: We are actually not finding many repeat mutations except in some loci which are known to be relatively hypermutable, such as the *Steel* locus.

Goodfellow: But these numbers are nowhere near saturation.

Brown: It is difficult to know that, because we don't really know what the underlying rate of recessive versus dominant mutation frequencies are, but our guess is that 40 000 ought to be approaching saturation.

Rubin: It will depend significantly on whether or not you are dealing with haploinsufficiency.

Goodfellow: Drosophila experiments indicate that you will not reach saturation with this number of mice.

Rubin: We have done screens and looked at a million genomes, and I wouldn't call that saturating. Also, the mutagens we use give a certain amount of specificity.

Brown: That is correct: the ENU mutagen works preferentially on AT base pairs.

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

Biological annotation of the *Drosophila* genome sequence

Gerald M. Rubin

University of California at Berkeley, Howard Hughes Medical Institute, Berkeley CA 94720-3200, USA

The fruitfly *Drosophila* has been a major organism for biological research for nearly 90 years. What will be the major near-term contribution of model organisms such as *Drosophila melanogaster* to the understanding of human biology and medicine, and how will the information from the genome projects help? While the human genome contains approximately 60 000 genes, these genes will encode the components of perhaps only a few hundred multicomponent, core biological processes. Data from a large number of studies have shown that many of the components of these biological processes and the way in which they interact with each other will be conserved between the invertebrate model organisms and human. More surprising is the extent to which the developmental and physiological functions of these core processes appear to be conserved. The importance of invertebrate model organisms for medical research derives from the fact that the experimental tools exist in these model organisms, but not in humans, for assembling genes into pathways. Many of these issues have been discussed in more detail in Miklos & Rubin (1996).

The nucleotide sequence of the *D. melanogaster* genome will soon be available. The value of these sequence data will be enormously enhanced if the structure of each transcription unit and the functions of its protein products can be established. Gene sequence and expression pattern databases will be extremely powerful tools. However, the function of a protein in a multicellular organism depends on context and will almost certainly need to be determined by experimental analysis.

Neither the intellectual framework nor experimental tools for analysing complex gene networks are currently in place. There is reason for cautious

optimism that the complete genomic sequence of organisms will enable the necessary global approaches to study gene function and regulation. The conservation of gene structure and function during evolution will allow for the linking and sharing of information garnered in different experimental systems. But what data should be collected and how to interpret these data are much less clear.

Genetic screens for loss-of-function mutations that affect a particular process have and will continue to play an important role in understanding the function of genes. Such screens have been carried out for decades in *Drosophila*. With the continual incorporation of more clever and sophisticated phenotypic analyses this experimental approach has been applied to an increasingly wide range of developmental, physiological and behavioural processes. These studies share a lot in common with modern genome research in that they are wide in scope—all the genes in the genome are being assayed in a single experiment—and they are usually not intended to test a specific hypothesis. Such genetic approaches have proven to be very powerful in grouping genes together in pathways and in allowing an unbiased—or, ignorance-driven—attack on a problem. To facilitate such studies, the Berkeley *Drosophila* Genome Project (BDGP) is carrying out gene disruption projects, using transposable element-mediated insertional mutagenesis, of unprecedented scale in a metazoan organism. To date over one-quarter of all essential genes have been mutated (Spradling et al 1995, 1999). Mapping the location of *P* element insertions in the BDGP strain collection relative to the 5′ ends of cDNAs and open reading frames observed in the genomic DNA sequence provides a powerful means of linking genes and phenotypes. These gene disruption experiments are now being extended to include transposable elements that can cause controlled misexpression of the gene at the site of insertion (see Rørth et al 1998).

However, these approaches have many inherent limitations. Genetics is an abstract science and its true power is only realized when combined with biochemistry. Moreover, it is becoming increasingly clear that few if any simple linear pathways exist and that one must learn to deal with complex, dynamic networks of interacting gene products. These networks are highly resilient; disruption in only one in three genes has an obvious phenotype in yeast, worms, flies or mice. Of the 14 000 genes thought to

exist in *Drosophila*, only 4000 are thought to mutate to recognizable lethal, sterile, visible or behavioural phenotypes. Even when a phenotype is observed it reflects only that part of a gene's function that cannot be compensated for, rather than revealing the complete role of the gene in development and physiology.

In an attempt to better understand the power and limitations of current methods to annotate a *Drosophila* genomic sequence with features of biological interest—as well as to get a glimpse of the detailed organization of the *Drosophila* genome—we carried out an analysis of a contiguous sequence of nearly 3 Mb (Ashburner et al 1999). Because this region has been genetically characterized to a greater degree than any other comparable region in any metazoan, it offered an unparalleled opportunity to correlate a sequence and genetic analysis. A computational analysis of the sequence predicts 218 protein coding genes, 11 tRNAs and 17 transposable element sequences. At least 38 of the protein coding genes are arranged in clusters of from 2–6 closely related genes, suggesting extensive tandem duplication. The gene density is one protein coding gene every 13 kb; the transposable element density is one element every 171 kb. Over 650 chromosome aberration breakpoints map to this chromosome region; their non-random distribution on the genetic map reflects variation in gene spacing on the DNA. Of 73 genes in this region identified by genetic analysis, 49 have been located on the sequence; *P* element insertions have been mapped to 43 genes. Ninety-five (44%) of the known and predicted genes match a *Drosophila* expressed sequence tag (EST), and 144 (66%) have clear similarities to proteins in other organisms. Perhaps the most interesting results from this study came from comparing the properties of the genes with and without observable phenotypes. Genes known to have mutant phenotypes are more likely to be represented in cDNA libraries, and are far more likely to have products similar to proteins of other organisms, than are genes with no known mutant phenotype.

References

Ashburner M, Misra S, Roote J et al 1999 An exploration of the sequence of a 2.9 megabase region of the genome of *Drosophila melanogaster*—the '*Adh*' region. Genetics 153:179–219

Miklos GLG, Rubin GM 1996 The role of the genome project in determining gene function: insights from model organisms. Cell 86:521–529

Rørth PK, Szabo A, Bailey T et al 1998 Systematic gain-of-function genetics in *Drosophila*. Development 125:1049–1057

Spradling AC, Stern DM, Kiss I, Roote J, Laverty T, Rubin GM 1995 Gene disruptions using *P* transposable elements: an integral component of the *Drosophila* genome project. Proc Natl Acad Sci USA 92:10824–10830

Spradling AC, Stern D, Beaton A et al 1999 The BDGP gene disruption project: single *P*-element insertions mutating 25% of vital *Drosophila* autosomal genes. Genetics 153:135–177

DISCUSSION

Fraser: Is there a plan in place for going forwards with functional genomics in *Drosophila*?

Rubin: We will sequence cDNAs, and do large scale *P* element mutagenesis. For expression patterns our plan is to generate the probes to assay expression in embryos, which is easy, because you can fix embryos and process them in 96-well plates. We will then enlist the rest of the community by saying to people that we have all these probes that we know work in microtitre plates, and if you are willing to take a sample of these probes and hybridize them to brain sections or whatever your favourite tissue is, we'll give you the probes for free as long as you give us back the data for the community database. I've already had many volunteers to do this.

Cohen: Would they also be free for pharmaceutical companies willing to do the same?

Rubin: We are trying to build a database where the data is freely accessible to all workers. You wouldn't be discriminated against if you were a pharmaceutical company, but you would have to be willing to give back the data. Again, I think this is in the interest of pharmaceutical companies or anyone who wants to use *Drosophila* sequences, because all the contributors are adding value to the sequence information and then that data is being made freely available.

Venter: It seems that *Drosophila* may soon be the best annotated genome, for the reasons that you have described. We have been talking about doing a *Drosophila*/*Caenorhabditis elegans* comparison. It is tough to do. If *C. elegans* is 30% over annotated, you are comparing noise to noise. What we need is well-annotated genes so we know what it is that we are actually comparing.

Rubin: Because the community is large, there are slightly over 2200 different *Drosophila* genes that were put in Genbank as individual genes

by people who wrote papers about that one gene. 1100 of those actually have a mutant phenotype that has been characterized. This is a much larger number of well-characterized genes than exists in *C. elegans*, even though that genome has been completely sequenced. This is because there are five to 10 times more people working on *Drosophila*.

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

Applications of high-density oligonucleotide arrays

Robert J. Lipshutz

Affymetrix Inc., 3380 Central Expressway, Santa Clara, CA 95051, USA

The Human Genome Project and other parallel commercial efforts are providing pharmaceutical researchers with access to unprecedented amounts of raw genomic sequence information, a nearly complete catalogue of human genes and a growing catalogue of human genetic variation. To effectively harness this information and apply it to pharmaceutical discovery, development, clinical trials and patient management, powerful new tools for measuring gene expression, discovering polymorphism and genotyping known variants are needed. GeneChip® high-density arrays of oligonucleotide probes are powerful tools to meet these requirements.

High-density arrays of oligonucleotide probes are synthesized by a unique combination of photolithography and solid phase chemical synthesis (Fodor et al 1991, 1993, Fodor 1997, McGall et al 1997, Pease et al 1994, Pirrung et al 1998, Southern et al 1992). This powerful approach allows large-scale parallel synthesis of thousands of compounds simultaneously in a miniaturized combinatorial format. Highly efficient strategies can be used to synthesize arbitrary polynucleotides at specified locations on the array in a minimum number of chemical steps (Fodor et al 1991). For example, the complete set of 4^N polydeoxynucleotides of length N, or any subset, can be synthesized in only $4 \times N$ cycles. Thus, given a reference sequence, a DNA probe array can be designed that consists of a highly dense collection of complementary probes with virtually no constraints on design parameters. The amount of nucleic acid information encoded on the array in the form of different probes is limited only by the physical size of the array and the achievable lithographic resolution. Current commercial bulk manufacturing methods allow for $\sim$409 600

polydeoxynucleotides to be synthesized on small 1.28×1.28 cm arrays. Experimental versions now exceed one million probes per array.

We have built an integrated system around the arrays including an easy to use polymeric cartridge with an integrated hybridization flow chamber, a fluidic station to control array hybridization, washing and staining, a confocal fluorescent scanner to collect data, and a fully integrated data storage, management and analysis system.

Once sequence information (partial or complete) for a gene is obtained, the next question is generally, 'What does the encoded protein do?' To understand function it is important to know when and where a gene is expressed, and under what circumstances the expression level is affected. Beyond questions of individual gene function are also questions concerning functional pathways and how cellular components (proteins as well as other molecules) work together to regulate and carry out cellular processes.

Addressing these questions requires the quantitative monitoring of the expression levels of very large numbers of genes repeatedly, routinely and reproducibly, while starting with a reasonable number of cells from a variety of sources and under the influences of genetic, biochemical and chemical perturbations. High-density oligonucleotide arrays have been shown to be very well suited for this task (Lockhart et al 1996, Mack et al 1998, de Saizieu et al 1998), allowing the simultaneous monitoring of all yeast genes (Cho et al 1998, Gray et al 1998, Wodicka et al 1997), all *Escherichia coli* genes, tens of thousands of human and mouse genes, and selective subsets of genes from a wide range of organisms.

The current performance limits of these tools have been determined and are described in Table 1 and Fig. 1.

Variation in DNA sequence underlies most of the differences we observe within and between species. Locating, identifying and cataloguing these genotypic differences are the first steps in relating genetic variation to phenotypic variation in both normal and diseased states. The use of high-density oligonucleotide arrays for genetic analysis has already proven to be very powerful. The design of arrays for this purpose is straightforward. Given a reference sequence for a region of DNA, four probes are designed to interrogate a single position. The set of four probes are typically centred at the interrogation position and one of them is designed to be perfectly complementary to a

TABLE 1 Gene expression oligonucleotide array performance characteristics

	Routine use	*Current limit*
Starting material[a]	5 μg total RNA	0.5 μg total RNA
Detection specificity[b]	1:100 000	$1{:}2\times10^{-6}$
Difference detection	Twofold changes	10% changes
Absolute quantitative accuracy[c]	$\pm2\times$	$\pm2\times$
False positives[d]	$<2\%$	$<0.1\%$
Discrimination of related genes[e]	70–80% identity	93% identity
Dynamic range (linear detection)[f]	$\sim$500-fold	$\sim10^4$-fold
Number of probe pairs per gene or EST[g]	20	4
Number of genes per array	7000	40 000

Performance characteristics for eukaryotic expression experiments using sets of 20 probe pairs per gene or expressed sequence tag (EST), 24 micron synthesis features (more than 280 000 features per 1.28×1.28 cm array), overnight hybridizations of biotin-labelled, randomly fragmented cRNA, and standard washing, staining, detection and image analysis protocols. The typical time required for a high-resolution (3 micron pixels) fluorescence scan is less than 10 min. Labelled samples are typically hybridized to arrays between two and 10 times without significant loss of performance (arrays are used for a single hybridization only).

[a]Total RNA is used directly without poly(A)$^+$ pre-purification steps. mRNA is converted to cDNA using a dT-primed reverse transcription reaction. The cDNA is made double-stranded and then transcribed into cRNA in an *in vitro* transcription (IVT) reaction. The IVT reaction results in a linear, unbiased amplification (typically 30- to 100-fold) of the original mRNA population (Gingeras et al 1998, Mack et al 1998).

[b]Results obtained using recommended post-hybridization signal amplification protocols. Detection of spiked RNAs at a relative abundance of less than $1{:}10^6$ has been achieved for a variety of transcripts in the presence of both human and mouse complex RNA samples (H. Dong & D. J. Lockhart, unpublished results).

[c]The hybridization signal intensities (PM minus MM values averaged over the probe pairs in the set) have been shown to be directly proportional to RNA concentration, and are predictive of absolute RNA concentration within a factor of two (Gingeras et al 1998, Mack et al 1998).

[d]False positives are defined on the basis of experiments in which samples are split, hybridized to different arrays, and the results compared (done with a wide range of human, mouse and yeast mRNA samples and arrays). A false positive is indicated if a probe set is scored quantitatively as an 'Increase' or 'Decrease' (on the basis of an analysis of the overall patterns) and quantitatively as changing by at least twofold. The extremely low false positive rates of less than 0.1 % are obtained using arrays synthesized on the same wafer and using simple repeated array scans and multiple-image data analysis methods.

[e]Probes are chosen from regions of sequence that are most different between family members, when known. Because of the targeted design of short oligonucleotides and the use of multiple probes per gene, it is possible to distinguish between very closely related sequences. For example, the yeast histone genes HTA1 and HTA2 are 93% identical at the DNA level (98% at the amino acid level). It was possible to design more than 10 25-mer oligonucleotides that were sufficiently different between the two sequences to allow unambiguous, independent detection of the two RNAs (L. Wodicka & D. J. Lockhart, unpublished results). The histone genes HTB1 and HTB2 are also highly similar (87% identical at the DNA level) but were independently detected as well.

[f]The range of RNA abundance over which hybridization signal intensity is linearly related to concentration (linearity within approximately a factor of two at the extremes of concentration). The extended dynamic range is achieved by combining the results of repeated scans at different wavelengths and/or detection system gains.

[g]Most expression arrays currently use between 15 and 20 probe pairs per gene or EST to increase sensitivity and quantitative accuracy and reduce the rate of false calls. Probe sets containing as few as four probe pairs (chosen using standard a priori probe design methods and without direct empirical data) have been used to semi-quantitatively screen very large numbers of ESTs for expression changes.

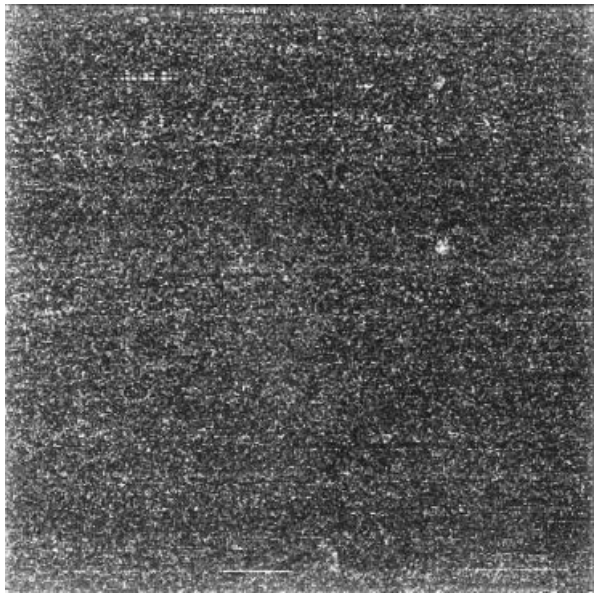


FIG. 1. Gene expression monitoring with oligonucleotide arrays. A single 1.28×1.28 cm array containing probe sets for approximately 40 000 human genes and expressed sequence tags.

short stretch of the reference sequence. The other three are identical to the first except at the interrogation position where the other three possible bases are substituted (Fig. 2). In the presence of a sample corresponding to the reference sequence, the probe complementary to the reference sequence will generally have the highest fluorescence intensity. In the presence of a sample with a different base at the interrogation position (a substitution variant), the probe corresponding to the variant base will have the highest fluorescence intensity. To interrogate one thousand bases of sequence, 1000 sets of four probes are used for highly parallel, comparative hybridization measurements. Similarly, if a specific single nucleotide polymorphism (SNP) is known, then tilings corresponding to each of the alleles can be encoded on the array providing a powerful assay for homozygous and heterozygous samples. These tools have been successfully applied to a large scale survey of SNPs in sequence-tagged sites (Wang et al 1998), the human mitochondrial sequence (Chee et al 1996), HIV-1 (Kozal et al 1996), cystic fibrosis (Cronin et al 1996) and other genes of interest.

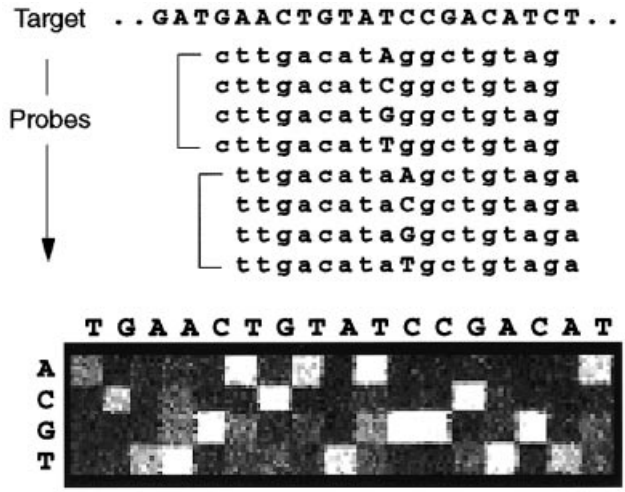


FIG. 2. Sequence analysis arrays: general tiling strategy. Detection of mutations or polymorphisms in a sequence is accomplished by using a four-probe interrogation strategy. In this illustration, four 17-mer oligonucleotide probes are used to determine the identity of the base in the middle of the probe sequence. The probe that forms the most stable duplex will provide the highest fluorescent signal among the four probes assigned to interrogate the central base. The next nucleotide in the target sequence is interrogated in the same manner using another set of four oligonucleotide probes. Probes with interrogation positions other than the central position, or probes of different lengths, can also be used to query the targeted base. Analysis of both strands of a target can be carried out on the same array to increase the confidence of the base determination.

Once an array is designed based on a set of reference sequences, the hybridization patterns can be used to classify samples into groups even without a determination of the exact sequence. In an application of this pattern-based approach, Gingeras et al (1998) and Troesch et al (1999) used high-density oligonucleotide arrays designed relative to the beta subunit of the RNA polymerase (rpoβ) and 16S ribosomal genes in *Mycobacterium tuberculosis* to accurately classify unknown samples from different species of the *Mycobacterium* genus.

High-density DNA probe arrays are powerful tools for a broad set of applications including gene expression monitoring, sequence analysis and genotyping. As the feature size shrinks, the information capacity of the array increases (Table 2). With recent adaptation of semiconductor-like photoresist processes, Beecher et al (1998) demonstrated the ability to

TABLE 2 Array capacity and feature size

Feature size	*Expression*[a]	*Sequence analysis*[b]	*Genotyping*[c]
50 μm	1600–6400 genes	8–16 kbp	2000–4000 markers
20 μm	10 000–50 000 genes	50–100 kbp	12 000–25 000 markers
2 μm	>1 million genes	500–1000 kbp	$1.2–5.5\times10^6$ markers

All numbers calculated for 1.28×1.28 cm arrays.
[a]Assuming 4–20 probe pairs per gene.
[b]Assuming 4–8 probes per base pair.
[c]Assuming 6–32 probes per marker.

synthesize arrays with features as small as two microns. At 2 μm resolution, one hundred million non-overlapping 30-mer probes spanning the entire human genome would fit on a 2×2 cm array.

References

Beecher JE, McGall GH, Goldberg MJ 1998 Chemically amplified photolithography for the fabrication of high density oligonucleotide arrays. Polym Mater Sci Eng 76:597–598

Chee M, Yang R, Hubbell E et al 1996 Accessing genetic information with high-density DNA arrays. Science 274:610–614

Cho RJ, Campbell MJ, Winzeler EA et al 1998 A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 2:65–73

Cronin MT, Fucini RV, Kim SM, Masino RS, Wespi RM, Miyada CG 1996 Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. Hum Mutat 7:244–255

de Saizieu A, Certa U, Warrington J, Gray C, Keck W, Mous J 1998 Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays. Nat Biotechnol 16:45–48

Fodor SPA 1997 DNA sequencing: massively parallel genomics. Science 277:393–395

Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D 1991 Light-directed, spatially addressable parallel chemical synthesis. Science 251:767–773

Fodor SPA, Rava RP, Huang XC, Pease AC, Holmes CP, Adams CL 1993 Multiplexed biochemical assays with biological chips. Nature 364:555–556

Gingeras TR, Ghandour G, Wang E et al 1998 Simultaneous genotyping and species identification using hybridization pattern recognition analysis of generic *Mycobacterium* DNA arrays. Genome Res 8:435–448

Gray NS, Wodicka L, Thunnissen A-M et al 1998 Exploiting chemical libraries structure, and genomics in the search for kinase inhibitors. Science 281:533–538

Kozal M, Shah N, Shen N et al 1996 Extensive polymorphisms observed in HIV-1 clade B protease gene using high density oligonucleotide arrays: implications for therapy. Nat Med 7:753–759

Lockhart DJ, Dong H, Byrne MC et al 1996 Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14:1675–1680

Mack DH, Tom EY, Mahader M et al 1998 Deciphering molecular circuitry using high-density DNA arrays. In: Mihich E, Croce C (eds) Biology of tumors. Plenum Press, New York, p 85–108

McGall GH, Barone AD, Diggelmann M, Fodor SPA, Gentalen E, Ngo N 1997 The efficiency of light-directed synthesis of DNA arrays on glass substrates. J Am Chem Soc 119:5081–5090

Pease AC, Solas D, Sullivan EJ, Cronin MT, Homes CP, Fodor SPA 1994 Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc Natl Acad Sci USA 91: 5022–5026

Pirrung MC, Fallon L, McGall G 1998 Proofing of photolithographic DNA synthesis with 3′,5′-dimethoxybenzoinyloxycarbonyl-protected deoxynucleoside phosphoramidites. J Org Chem 63:241–246

Southern E, Maskos U, Elder JK 1992 Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models. Genomics 13:1008–1017

Troesch A, Nguyen H, Miyada CG et al 1999 Mycobacterium species identification and rifampin resistance testing with high-density DNA probe arrays. J Clin Microbiol 37:49–55

Wang DG, Fan J-B, Siao C-J 1998 Large-scale identification, mapping, and genotyping of single nucleotide polymorphisms in the human genome. Science 280:1077–1082

Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ 1997 Genome-wide expression monitoring in *Saccharomyces cerevisiae*. Nat Biotechnol 15:1359–1367

DISCUSSION

Roses: In genotyping, is the 1.5% error in the same SNP assays, or is it a random 1.5% in multiple assays?

Lipshutz: That's a random 1.5%: it is noise. In an assay like this, you can toss out poorly performing markers, so anything that is systematically bad we just remove from the assay beforehand.

Venter: Does that imply that the reproducibility is improving? There were complaints about the chip-to-chip variability with your HIV test and also that there were problems with the degradation of the synthesis on the chips.

Lipshutz: On the arrays we did 10 samples four times, and looked at 1494 markers. The failure to reproduce was at a level of 0.38%

Venter: What happens when you do it in 'untrained' labs? Does that change?

Lipshutz: We are continuing to explore that. What I'm saying is it is not the array that is the limiting factor, it is basically the robustness of the sample prep.

Goodfellow: I'm still confused about that 1.5% genotyping error. If your within-sample reproducibility is 0.3%, what does that does that 1.5% represent?

Lipshutz: Those two statistics were generated on different samples. There could be samples that are reproducibly wrong. Saying that it is 'reproducible' says that you get the same answer each time. We also said

that across a selection of samples, in 1.5% the answer was different doing by the array than doing it by gel-based sequencing.

Goodfellow: So if you throw out that 1.5%, you will then find that the error rate diminishes.

Venter: The consistency rate, or the error rate?

Goodfellow: That's what I'm pushing, because I'm not quite sure I understand what it is that we are discussing here.

Lipshutz: The concordance rate with gel-based sequencing was determined in the following way. We took about 700 of the markers and 44 samples, and we did double-stranded gel-based sequencing on all of those. We then tried to interpret as best we could and with expert help the genotypes of each one of the markers. We then independently computed the genotypes based on the algorithms that we had already established for the arrays. When we compared those, if you look at the total number (which is something like 700 markers times 40 samples), in 1.5% of the genotypes there was a difference between the results.

Goodfellow: But it then depends on how that 1.5% was distributed.

Lipshutz: It is random. This is basically the system noise between either the gels or the arrays.

Hochstrasser: How far are you from clinical applications?

Lipshutz: We have three array-based assays that are sold as analyte-specific reagents, which is the Food and Drug Administration (FDA) requirement for using something in what is called a homebrew test. Those three arrays have been incorporated into CLIA (Clinical Laboratory Information Act) approved assays by reference laboratories in the USA. This is for HIV, P450 and p53. Also, in our collaboration with bioMérieux we are preparing to take a tuberculosis classification and drug resistance test to the clinic and do FDA testing over the next few years.

Venter: When you compare the Affymetrix yeast chip to the Stanford gene array chips, do you get the same answers?

Winzeler: I've been hearing that you generally get the same answer but the Affymetrix system is more sensitive for low-level transcripts. For the abundant transcripts you can detect the same sorts of differences that you can detect with microarray hybridizations, but it depends on the experiment and whether you do multiple hybridizations.

Venter: The big question in the field is whether the answers you get are robust or not. With so many data points on there (and The Institute for Genome Research was really struggling with this), how do you ensure that each point is a meaningful datum and not just noise? There are different rumours out there for each data set. It is hard to pin down in any quantitative way.

Lipshutz: Our customers have done a lot of Northern validations. We have done extensive testing and characterization of our assays (Lipshutz et al 1999). Furthermore, many independent users have done their own comparisons with very satisfactory results (Li et al 1999, Alon et al 1999, Jelinsky & Samson 1999, Famborough et al 1999, Wang et al 1999, Harkin et al 1999, Zhu et al 1998, Der et al 1998, Holstege et al 1998, Cho et al 1998, Gray et al 1998, Wodicka et al 1997, Lockhart et al 1996). In addition, several investigators are currently doing side-by-side tests with other technologies and we expect they will also publish.

Goodfellow: If you want to do a comparison, the last thing you do is a Northern blot. If there was ever an assay which is non-quantitative, it is the Northern blot.

Fraser: I have a question about the leukaemia study. How many human genes were screened, and were they selected at random? And how many were informative in terms of changes in expression levels?

Lipshutz: They used an array with about 6500 full length human genes. It is a standard commercial array. A ranking system was used to order them as to informativity based on how well the distributions of expression patterns were separated for the two different classes. Up to about 200–300 of the different genes actually provided statistically significant differences that could be used. The algorithm was tested with anywhere from about 10 to 200 genes. In each case it performed at a similar level. We don't know how many genes we would eventually want to use in the test, but I believe that if you use a larger number (closer to 50) you would be less subject to noise in any one particular gene.

Fraser: Do you think the results would change if this had been a sample of solid tumours rather than leukaemias, where, depending upon surgical margins, you may be looking at both normal and tumour samples?

Lipshutz: That's another good question. Investigators at the Whitehead Institute are doing some of those tests now, and the preliminary results look fairly positive, but I don't have the data.

Hochstrasser: You should use laser dissection microscopy to get an adequate sample of pure tissue.

Lipshutz: There is a trade-off there. One of the hopes is that if you look at a large enough set of genes, there will be genes that are uniquely expressed in the cancer. In that way you can get through some of that tissue variation.

References

Alon U, Barkai N, Notterman DA et al 1999 Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 96:6745–6750

Cho RJ, Campbell MJ, Winzeler EA et al 1998 A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 2:65–73

Der SD, Zhou A,Williams BRG, Silverman RH 1998 Identification of genes differentially regulated by interferon alpha, beta, or gamma using oligonucleotide arrays. Proc Natl Acad Sci USA 95:15623–15628

Famborough D, McClure K, Kazlauskas A, Lander ES 1999 Diverse signaling pathways activated by growth factor receptors induce broadly overlapping, rather than independent, sets of genes. Cell 97:727–741

Gray NS, Wodicka L, Thunnissen AW et al 1998 Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. Science 281:533–537

Harkin DP, Bean JM, Miklos D et al 1999 Induction of GADD45 and JNK/SAPK-dependent apoptosis following inducible expression of BRCA1. Cell 97:575–586

Holstege FCP, Jennings EG, Wyrick JJ et al 1998 Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95:717–728

Jelinsky SA, Samson L 1999 Global response of *Saccharomyces cerevisiae* to an alkylating agent. Proc Natl Acad Sci USA 95:1486–1491

Li H, Schrick JJ, Fewell GD et al 1999 Novel strategy yields candidate Gsh-1 homeobox gene target using hypothalamus progenitor cell lines. Dev Biol 211:64–76

Lipshutz RJ, Fodor SPA, Gingeras TR, Lockhart DJ 1999 High density synthetic oligonucleotide arrays. Nat Genet (suppl) 21:20–24

Lockhart DJ, Dong H, Byrne MC et al 1996 Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14:1675–1680

Wang Y, Rea T, Bian J, Gray S, Sun Y 1999 Identification of the genes responsive to etoposide-induced apoptosis: application of DNA chip technology. FEBS Letts 445:269–273

Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ 1997 Genome-wide expression monitoring in *Saccharomyces cerevisiae*. Nat Biotechnol 15:1359–1367

Zhu H, Cong JP, Mamtora G, Gingeras T, Shenk T 1998 Cellular gene expression altered by human cytomegalovirus: global monitoring with oligonucleotide arrays. Proc Natl Acad Sci USA 95:14470–14475

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

Plasmodium falciparum: from genomic sequence to vaccines and drugs

Stephen L. Hoffman and Daniel J. Carucci

Malaria Program, Naval Medical Research Center, 503 Robert Grant Avenue, Silver Spring, MD 20910, USA

Malaria parasites infect 300–500 million and kill 1.5–2.7 million people annually. In areas with intense transmission, each infected individual may harbour more than five different strains of *Plasmodium falciparum*. There is now a major international effort to sequence the 30 megabase genome of *P. falciparum*. Initially there was scepticism within the field that the genome could be sequenced completely. This was in large part due to the instability of the DNA in *Escherichia coli*, a phenomenon thought to be at least in part due to the high (80.2%) A+T content of the genome. With the publication of the 947 kb sequence of chromosome 2 of *P. falciparum* (Gardner et al 1998), the capacity to sequence the genome has been established, and there is great hope that elucidation of the sequence of the estimated 6000 genes on 14 chromosomes will lead to increased understanding of the biology of the parasite and then to the development of new drugs to treat and prevent the disease and vaccines that will prevent the development of disease and death.

The question is how will this be accomplished? The two most important groups of drugs for the treatment of malaria are those based on quinine and artemesinin. Quinine is derived from the bark of the cinchona tree and has been in use for at least 350 years, since being brought by Jesuits from Peru to Europe. Artemesinin is derived from *Artemesia annua* (sweet wormwood) and has been used to treat malaria for more than 2000 years in China where it is known as Qinghaosu. Beginning in the late 1950s with chloroquine, perhaps the best antimalarial ever developed, *P. falciparum*, the most important malaria-causing parasite, has demonstrated the capacity to develop resistance to

essentially all new antimalarials. There is currently no licensed antimalarial vaccine, and during the past 15 years 95% of all clinical trials of experimental malaria vaccines have tested immunogens based on sequences from only two *P. falciparum* proteins, the circumsporozoite protein (CSP) and the major merozoite surface protein (MSP1) (Miller & Hoffman 1998). Based on this sparse record of accomplishment in developing and sustaining new drugs, and testing experimental vaccines, many have been sceptical of our capacity to rationally and systematically utilize genomic sequence data to develop effective, sustainable new drugs and vaccines to control and then eliminate the parasite.

Developing such strategies requires understanding the life cycle of the parasite (Fig. 1). In contrast to viruses and bacteria, the parasites that cause malaria have a complex life cycle and many of the proteins expressed at one stage of the life cycle are not present at other stages. *Anopheles* sp. mosquitoes inoculate uni-nucleate sporozoites which rapidly enter the circulation, home to the liver and invade hepatocytes where they do not cause any pathology or disease, but develop during a week to mature liver stage schizonts with 10 000–40 000 nucleated 'merozoites'. The ideal chemoprophylactic drug or vaccine would prevent parasites from maturing within hepatocytes and escaping into the bloodstream, targeting molecules potentially unique to the sporozoite or liver stages. A drug for treating individuals ill with malaria would have to target proteins and other molecules present at the erythrocytic stage of the life cycle, the stage responsible for all clinical manifestations of the disease. Finally, a drug or vaccine designed exclusively to reduce transmission of the parasite to mosquitoes would have to target the sexual stage of the life cycle.

Thus, many believe that the first step in effectively utilizing genomic sequence data for drug and vaccine development is characterization of stage-specific expression of parasite genes and proteins. At the RNA level this could be accomplishing by creating DNA microarrays or DNA chips, extracting RNA from each stage of the life cycle, synthesizing cDNA and determining the relative expression of the different genes. The asexual and sexual erythrocytic stages of *P. falciparum* can be cultivated *in vitro* in erythrocytes, making this process relatively straightforward, and we have assessed the 209 protein-encoding genes from chromosome 2 of *P. falciparum* using a DNA

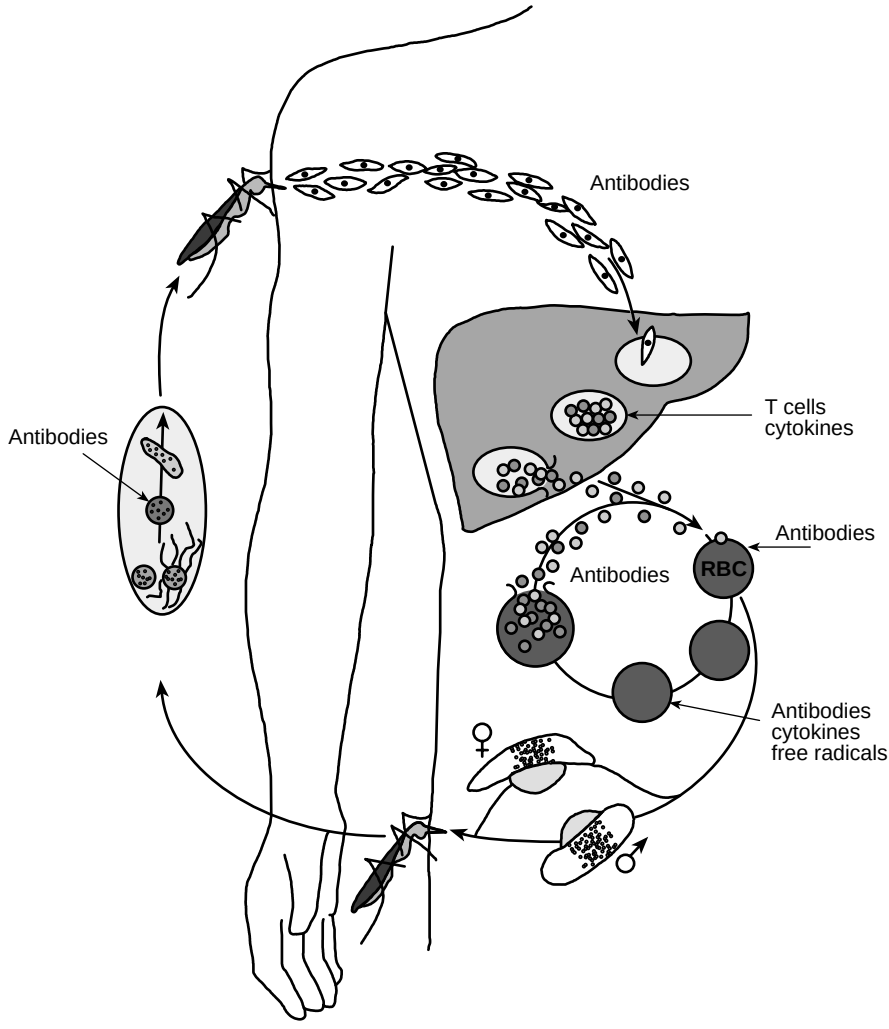


FIG. 1. Life cycle of *Plasmodium falciparum*.

microarray (Fig. 2). However, sporozoites can only be produced in mosquitoes, and liver stages within primary human hepatocytes. Producing enough material for screening microarrays and DNA chips is difficult (sporozoites) and essentially impossible (hepatic stage) without the use of currently unstandardized amplification techniques. Our approach to establishing stage-specific expression, particularly at the hepatic stage, is to focus on protein expression. We are constructing

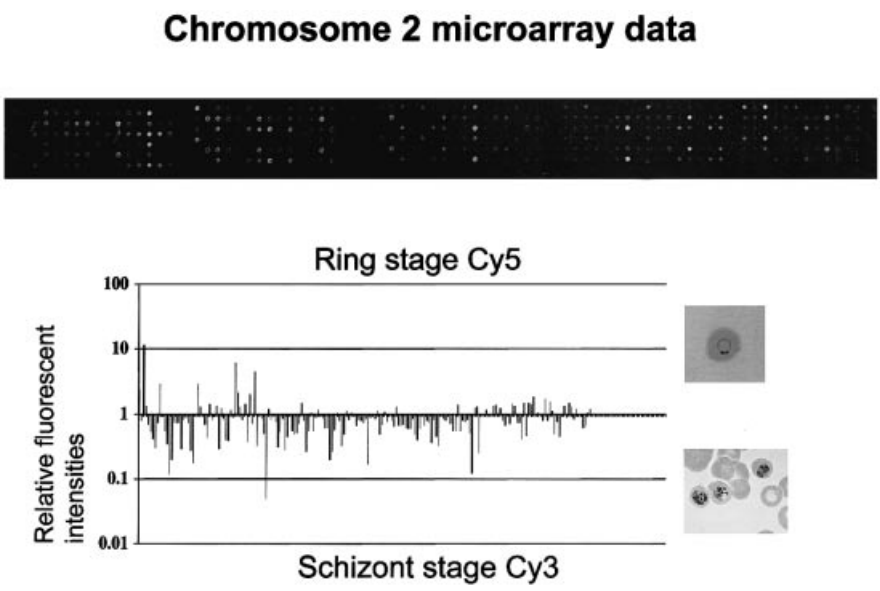


FIG. 2. Expression of the 209 genes on chromosome 2 of *P. falciparum* at the early ring erythrocytic stage, and the late schizont erythrocytic stage of the life cycle as determined by hybridization of cDNA from these stages against the 209 genes arrayed as a DNA microarray on a slide.

DNA vaccines for each open reading frame (the vaccinome), immunizing groups of mice with individual plasmids and then screening the antisera against the different stages of the parasite life cycle (Hoffman et al 1998). By establishing gene expression where possible, and protein expression by recognition of antibodies, we will begin to establish the stage-specific expression of the genes in the *P. falciparum* genome.

Having established stage-specific expression, identification of critical targets for drug and vaccine development is still complex. Annotation and bioinformatics are then used to predict trafficking and function of potential target proteins. Results from the analysis of the sequence of *P. falciparum* chromosome 2, comprising 3% of the genome, are shown in Table 1 and Fig. 3. With 6000 genes there will be many leads that need to be followed. It would be useful to be able to systematically knock out, or knock in genes in the genome to determine which are critical for survival, but it is not yet possible to do this rapidly for *P. falciparum*, although it can be done for individual genes of interest in a few laboratories (Ménard et al 1997).

TABLE 1 Predictions regarding protein encoding genes on chromosome 2 of *P. falciparum* (Gardner et al 1998)

Total protein encoding genes	209
Secreted	22 (11%)
Integral membrane	90 (43%)
Integral membrane with multiple transmembrane domains	27 (13%)
With multiple coiled coil domains	111 (53%)
With non-globular domains[a]	155 (75%)
Completely non-globular	17 (8%)

[a]The term 'non-globular' refers to proteins or domains of proteins that do not assume compact, folded structures.

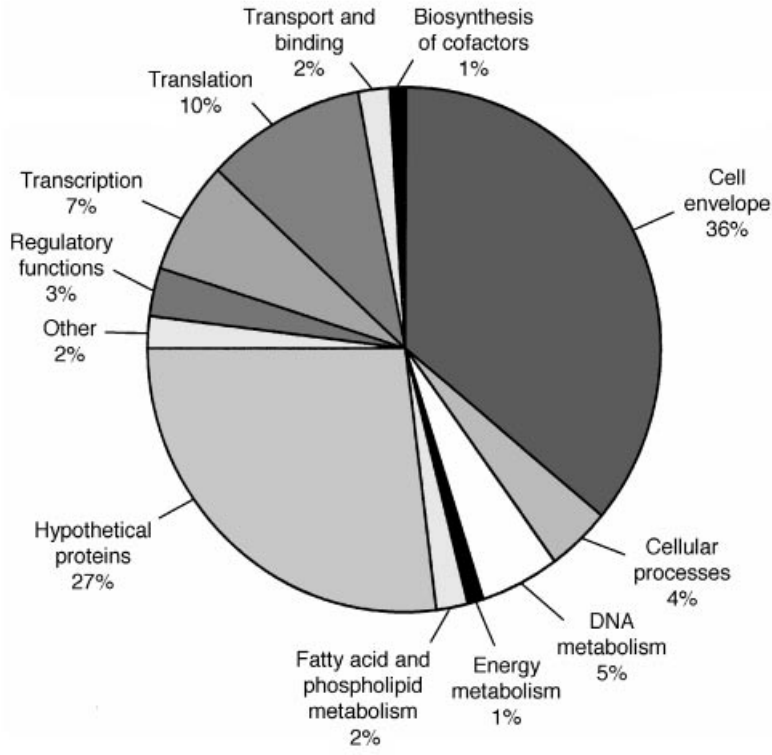


FIG. 3. Genes by role category from chromosome 2 of *P. falciparum* (Gardner et al 1998).

We are focusing on using the data for vaccine development. One approach is to develop a vaccine that prevents parasites from emerging from the liver into the bloodstream, thereby preventing all clinical manifestations of the disease. Immunization of volunteers with radiation-attenuated *P. falciparum* sporozoites protects them completely against challenge with live parasites for at least nine months. The protection is species, but not strain specific, and is thought to be primarily mediated by CD8+ T cells that recognize antigens expressed by irradiated sporozoites within hepatocytes (reviewed in Hoffman et al 1996). It is impractical to immunize large numbers of individuals by the bite of irradiated, *P. falciparum*-infected mosquitoes, so we have been working to develop a subunit vaccine that duplicates this immunity. One of our approaches is to identify all proteins expressed by irradiated sporozoites within hepatocytes as described above, predict all of the 8–10 amino acid peptides from these proteins that should bind to the three HLA class I superfamily molecules, HLA-A2, -A3/A11 and -B7, and the allelles HLA-A1 and -A24, synthesize those peptides, assess the peptides for degenerate binding to the respective members of these class I superfamilies or class I molecules, and determine whether volunteers immunized with *P. falciparum* sporozoites mount CD8+ T cell responses against these peptides (Doolan et al 1997). We will then synthesize genes that include all of these sequences, and construct DNA vaccines, and recombinant poxvirus vaccines that include these synthetic genes. A DNA vaccine prime, recombinant poxvirus boost immunization regimen then will be assessed for safety, immunogenicity and protective efficacy in volunteers (Sedegah et al 1998, Schneider et al 1998).

As we come closer to finishing the sequence of *P. falciparum*, it becomes increasingly critical to develop systematic, high throughput methods for identifying genes that encode proteins which are important targets for drug and vaccine development. The process will become more interesting, more complex, and perhaps more rewarding as more data emerge from the human genome project. In fact, the challenge of the 21st century will be to utilize data from the malaria and human genome projects to create new drugs and vaccines for controlling *P. falciparum* malaria, the single most important infectious disease cause of death of young children in the world.

Acknowledgements

This work was supported by Naval Medical Research and Development Command work units 61102A.S13.00101.BFX1431, 61278A.870.0010. EFX1432, 623002A.810.00101.HFX.1433 and STEP C611-102A0101BCX. The opinions and assertions herein are those of the authors and are not to be construed as official or as reflecting the views of the US Navy or naval service at large.

References

Doolan DL, Hoffman SL, Southwood S et al 1997 Degenerate cytotoxic T cell epitopes from *P. falciparum* restricted by multiple HLA-A and HLA-B supertype alleles. Immunity 7:97–112

Gardner MJ, Tettelin H, Carucci DJ et al 1998 Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. Science 282:1126–1132

Hoffman SL, Franke ED, Hollingdale MR, Druihle P 1996 Attacking the infected hepatocyte. In: Hoffman SL (ed) Malaria vaccine development: a multi-immune response approach. ASM Press, Washington, DC, p 35–75

Hoffman SL, Rogers WO, Carucci DJ, Venter JC 1998 From genomics to vaccines: malaria as a model system. Nat Med 4:1351–1353

Ménard R, Sultan AA, Cortes C et al 1997 Circumsporozoite protein is required for development of malaria sporozoites in mosquitoes. Nature 385:336–340

Miller LH, Hoffman SL 1998 Research towards vaccines against malaria. Nat Med (suppl) 4:520–524

Schneider J, Gilbert SC, Banchard TJ et al 1998 Enhanced immunogenicity for CD8+ T cell induction and complete protective efficacy of malaria DNA vaccination by boosting with modified vaccinia virus Ankara. Nat Med 4:397–402

Sedegah S, Jones TR, Kaur M et al 1998 Boosting with recombinant vaccinia increases immunogenicity and protective efficacy of malaria DNA vaccine. Proc Natl Acad Sci USA 95:7648–7653

DISCUSSION

Venter: For the children that live past the age of five, what are their antibodies working against?

Hoffman: A whole range of proteins. Vaccine development in most cases is focused on one, two or three major surface proteins, whereas immune responses in individuals within the community are directed against perhaps thousands of proteins. However, the level and function of these antibody and T cell responses are generally poor. We can do better. I don't believe that doing better against one or two proteins will be adequate.

Venter: Can you use the natural serum to select those epitopes and then just engineer back from that?

Hoffman: That would be part of the screen.

Goodfellow: Hasn't that been done with the bacteriophage expression approaches?

Hoffman: Not very well. If you purify immunoglobulin G (IgG) from an individual in West Africa, take it to Thailand and administer it to a child with multidrug resistant malaria, the IgG will knock down the parasitaemia by 99%. That IgG has been used by some to try to determine what proteins are recognized. The problem is that the IgG contains antibodies that recognize practically all proteins. If you were to do a 2D gel on blood-stage parasites with that IgG you would have thousands of spots. We think that to be protective the IgG must be acting against accessible proteins: essentially proteins on the surface of the merozoites or infected erythrocytes. If we identify which genes encode proteins that induce antibodies that recognize surface-expressed proteins, that will cut the number considerably. If the B cell epitopes are linear one could then easily identify them. It is actually much more difficult to deal with conformational antibody epitopes than it is with linear nine amino acid CD8+ T cell epitopes. In fact, many of the important epitopes are formed only when the protein folds in its native structure.

Venter: Is there any way to block the uptake into the liver in the first place, instead of trying to do something inside the liver?

Hoffman: About 15 years ago a number of scientists thought they were going to win a Nobel prize for doing that! The gene for the major sporozoite surface protein, the CSP, was cloned in 1984 (Dame et al 1984). It was found that monoclonal antibodies against this protein passively protected mice and monkeys against malaria. The problem is that the sporozoites probably enter liver cells within 5 min of inoculation by mosquitoes, so the antibodies have to be present at protective levels. If 100 sporozoites are inoculated and one sporozoite gets in and develops, those antibodies won't do anything against the next stage of the life cycle.

Fraser: A limited number of algorithms have been developed to try to predict T cell epitopes. Do you know what the state of the art is with these? Have they been tested in terms of the validity of the targets that they identify?

Hoffman: We've been working with a group at Epimmune Inc. in San Diego that is expert in this area. We studied the first four genes that we

know were expressed by irradiated sporozoites in hepatocytes and predicted all of the peptides from these proteins that bound to multiple members of the three class I HLA superfamilies, HLA A2, HLA 3/11 and HLA B7. There were over 300 of such peptides. Through a series of experiments this was reduced to 17 peptides. Dr Denise Doolan then demonstrated that individuals naturally exposed to malaria in Kenya and individuals experimentally immunized with irradiated sporozoites made CD8+ T cell responses against all 17 peptides (Doolan et al 1997). We know that the process works. Whether it can be built up to the level for screening the genome and actually work is another story.

Venter: I noticed that there was a kanamycin-resistance gene on one of your vectors. Has there ever been a concern about immunizing humans with an antibiotic resistance gene?

Hoffman: There is in terms of ampicillin, because it's a commonly used antibiotic, and there are concerns about induction of ampicillin resistance. Kanamycin is actually well accepted as a selective marker in the field of recombinant proteins and DNA vaccines because it is rarely used as an antibiotic.

Goodfellow: Is the hypothesis that there are a few key antigens that you need to make antibodies against, or is it that you are going to have to make antibodies to hundreds of antigens?

Hoffman: There are several schools of thought. One would hold that there are several key antigens, and all we need is antibodies against a few of these. A lot of work is going on in trying to develop a conformationally appropriate immunogen to induce antibodies against the major merozoite surface protein, a protein thought to be one of these. Another school of thought argues the following. There are at least one and perhaps two variant surface proteins which get to the surface of erythrocytes and mediate binding to endothelial cells in postcapillary venules and capillaries. Protective immunity involves developing antibody responses against many of the variants of these proteins. Right now, no conserved regions on the variant proteins has been identified that we can actually target in vaccine development, but a lot of the work is going on to try to develop methods for targeting these variant proteins. A third approach which we are pursuing is to prime the immune system against as many surface-accessible proteins as possible and then have infection itself boost the immune response to these. We believe that this immune

response will then limit the infection, and stop people dying from malaria.

Venter: Can you immunize the mosquitoes?

Hoffman: There are scientists working on that. There are those who are also trying to immunize people against mosquitoes.

Venter: With many of these genomes, it's not hard to develop good vaccines against the clonal variety. One biotech company produced a large number of *Haemophilus* surface proteins and made antibodies against them. They found against that strain they were all protected, but as soon as they went out into the clinical situation they found that *Haemophilus* has so much built-in variation in the antigenic sites that they didn't work. Experimental models with clonal sets tell you nothing about the real world.

Hoffman: Does anyone have any thoughts about how we might more intelligently get at stage-specific expression? At the sporozoite/liver stage where we can't get that much material, is there some other technology we could use to systematically establish stage-specific expression?

Goodfellow: Is this human-specific, so you won't get infection in the mouse?

Hoffman: You can't infect the mouse.

Goodfellow: You could try to use the model system where human hepatocytes are cultured *in vivo* in mice. If you can infect the human cells you might be able to devise a selection, perhaps based on fluorescence-activated cell sorting, for separating out the infected cells.

Is there anything you can do to manipulate the genome, to make a transgenic malaria parasite?

Hoffman: It can be done with great difficulty. Knockouts are being made, genes are being overexpressed, but it can't be done systematically.

Goodfellow: Can you do it in strains which people use in infections? Why not try to specifically start tagging genes with an epitope for which you have good mononclonal antibodies?

Venter: One problem is the lack of robust *in vitro* cultivation systems. For example, it was difficult to obtain enough purified chromosomal material to sequence the *P. falciparum* genome.

Goodfellow: Yes, but the trick is enrichment.

Hoffman: This would of course require engineering a parasite that directed a selectable target to the surface of hepatocytes.

Goodfellow: Perhaps you could try using a human promoter in the parasite. If this works well enough you could use a liver-specific promoter to express green fluorescent protein. If you get this into the parasite, it could be used as a basis for selection to pull out the infected cells.

References

Dame JB, Williams JL, McCutchan TF et al 1984 Structure of the gene encoding the immunodominant surface antigen on the sporozoite of the human malaria parasite *Plasmodium falciparum*. Science 225:593–599

Doolan DL, Hoffman SL, Southwood S et al 1997 Degenerate cytotoxic T cell epitopes from *P. falciparum* restricted by multiple HLA-A and HLA-B supertype alleles. Immunity 7:97–112

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

Functional analysis of the yeast genome by precise deletion and parallel phenotypic characterization

Elizabeth A. Winzeler*, Hong Liang, Daniel D. Shoemaker and R. W. Davis

Department of Biochemistry, Stanford University School of Medicine, Stanford, CA 94305-5307, USA

New genes are being identified at an unprecedented rate in the genome sequencing projects. The functions of the proteins encoded by many of these genes are unknown. Classically, the analysis of mutants lacking a gene has provided the best clue to the gene's function. However, both mutating specific genes and assessing the phenotypes of the resulting mutants are labour-intensive processes. Using the genome of the baker's yeast *Saccharomyces cerevisiae* as a model we have been exploring modifications of this traditional approach to accommodate the accelerated pace of new gene identification. As such we are participating in an effort by an international consortium of laboratories to systematically create deletions in all the genes in the yeast genome (Winzeler et al 1999). The goal will be to screen these yeast deletion strains for specific phenotypes in order to assign functions to proteins encoded by the deleted gene. In addition the strains will be useful to the yeast researcher as well as to any investigator whose protein of interest has a yeast homologue.

The precise deletion of yeast genes can now be efficiently accomplished using a PCR-mediated gene disruption strategy that exploits both the high rate of homologous recombination in yeast and the availability of yeast genome sequence (Lorenz et al 1995, Wach et al

*Present address: Genomics Institute of the Novartis Research Foundation, 3115 Merryfield Row, Suite 200, San Diego, CA 92121, USA.

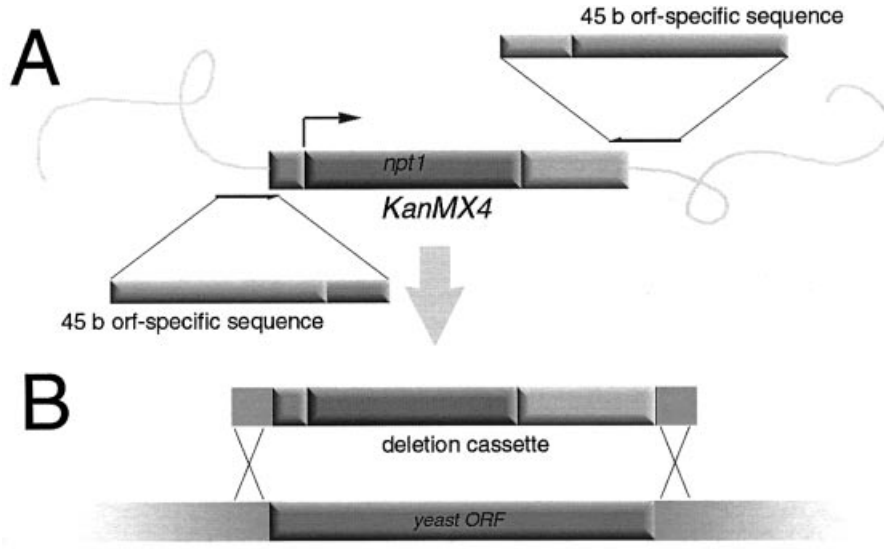
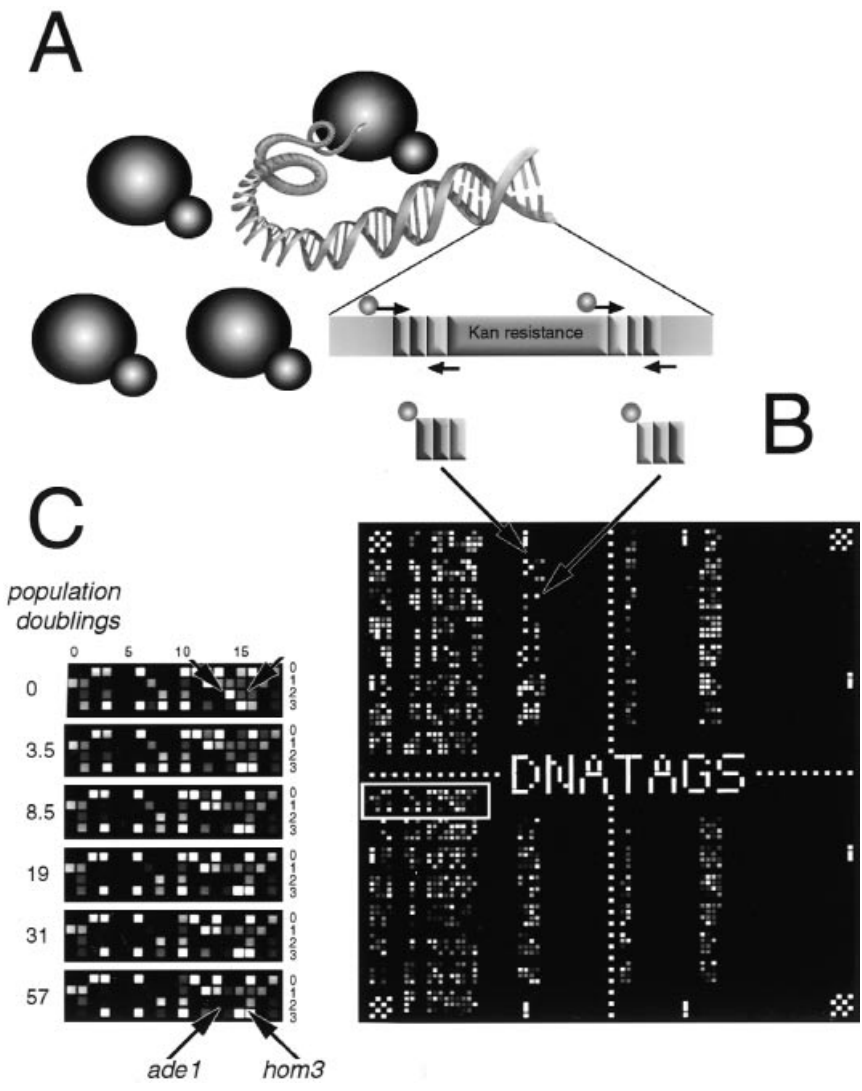


FIG. 1. Construction of deletion strains. (A) Long oligonucleotide primers are synthesized that contain yeast genome sequence on their 5′ ends and sequence complementary to a heterologous drug resistance cassette on their 3′ ends. The primers are used to PCR amplify the drug resistance module (KanMX4) (Wach et al 1994). (B) The resulting deletion cassette can be introduced into yeast where it will replace the targeted gene by homologous recombination under selection for drug resistance. ORF, open reading frame.

FIG. 2. (*opposite*) Parallel functional analysis of deletion strains. (A) In the process of deleting each gene each strain was marked by two molecular 'bar codes' that are specific to that deletion strain. The bar codes (essentially unique 20mers) are flanked by sequences that function as common PCR priming sites such that the bar codes can be collectively amplified from genomic DNA isolated from a heterogeneous pool of deletion strains using the same pairs of fluorescein or biotin-labelled primers. (B) Scanned image of a high-density oligonucleotide array carrying probes complementary to the bar codes carried in the deletion strains. The location of each bar code complement on the array is known. The array was hybridized with ~1100 biotin-labelled bar code amplicons. Not all bar code sequences hybridize to the array with the same affinity, but the hybridization behaviour is reproducible. (C) Negative selections using the tagged deletion strains. 558 homozygous diploid deletion strains were grown in minimal media for 57 population doublings. Aliquots of cells were collected from the pool at the indicated times and processed as described above. Portions of the scanned arrays indicated by the box in B are shown for the different time points. The hybridization intensity for bar codes from strains carrying deletions in all known auxotrophic genes in the pool, including the *adel* (position 14,2) and *hom3* (15,2) strains diminish with time, reflecting the loss of these strains from the pool. Hybridization intensities for bar codes (e.g. 8,3) from strains that are slow-growing in both rich and minimal media also decrease with time.

1994). In this method short regions of yeast sequence (45 base pairs) identical to those found upstream and downstream of the targeted gene are placed at each end of a selectable marker gene through the PCR (Fig. 1). For most genes, more than 95% of the resulting yeast transformants carry the correct deletion (Wach et al 1997). The main



advantage of this method is that it is highly automatable and no cloning steps are required. Using a high-throughput strategy more than 6000 different genes (approximately 95% of the predicted genes in the genome) have now been precisely deleted and verified by the different consortium laboratories (see *http://sequence-www.stanford.edu/group/yeast_deletion_project/deletions3.html* for a complete list of available strains).

In order to quicken the pace of deletion mutant characterization, the PCR-mediated disruption strategy was modified so as to introduce molecular 'bar codes' into the deletion strain. The bar codes are unique 20 base pair sequences that serve as strain identifiers. They are incorporated into the deletion strain by including the bar code sequences in the long oligonucleotide primers that are used to generate the deletion cassettes (Shoemaker et al 1996). These bar codes can be detected by hybridization to arrays of nucleic acids (Fig. 2B). Thus, they allow large numbers of deletion strains to be pooled and analysed in parallel in competitive growth assays (Fig. 2C). This direct, simultaneous, competitive measurement of fitness increases the sensitivity, accuracy and speed with which growth defects can be detected relative to conventional methods.

As an example, the phenotypes of the first 558 strains created in the deletion project were analysed collectively in a pool. The pool was grown in rich or minimal medium for many generations. Genomic DNA was isolated from the pool at different times and the abundance of particular bar codes, and hence the relative proportion of the corresponding deletion strain was measured. It was expected that known auxotrophic strains would disappear from the population when the pool was grown in minimal medium, but not when grown in rich medium (Fig. 2C). As predicted the bar code hybridization signals from the auxotrophic deletion strains, including the *ade1* and *hom3* strains, quickly became undetectable.

This ability to assess thousands of strains quantitatively and in parallel will significantly decrease the amount of labour and materials needed for drug-sensitivity screens (Giaever et al 1999), as well as increasing the reliability of the data interpretation and functional classifications. While the construction and verification of thousands of deletion strains requires a substantial investment of effort, once made, the strains will provide a

lasting resource. Their availability should substantially accelerate the process of assigning function to genome sequence.

References

Giaever G, Shoemaker D, Jones T et al 1999 Genomic profiling of drug sensitivities via induced haploinsufficiency. Nat Genet 21:278–283

Lorenz MC, Muir RS, Lim E, McElver J, Weber SC, Heitman J 1995 Gene disruption with PCR products in *Saccharomyces cerevisiae*. Gene 158:113–117

Shoemaker DD, Lashkari DA, Morris D, Mittmann M, Davis RW 1996 Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. Nat Genet 14:450–456

Wach A, Brachat A, Pöhlmann R, Philippsen P 1994 New heterologous modules for classical or PCR-based gene disruptions in *Saccharomyces cerevisiae*. Yeast 10:1793–1808

Wach A, Brachat A, Alberti-Segui C, Rebischung C, Philippsen P 1997 Heterologous HIS3 marker and GFP reporter modules for PCR-targeting in *Saccharomyces cerevisiae*. Yeast 13:1065–1075

Winzeler EA, Shoemaker D, Astromoff A et al 1999 Functional characterization of the *Saccharomyces cerevisiae* genome by gene deletion and parallel analysis. Science 285:901–906

DISCUSSION

Venter: Instead of doing those as a pool, if you took those strains out and grew them individually would you get the same answer, or would you get the same answer you got with gene expression? This is a problem we really worried about with the *Mycoplasma genitalium* deletions, where you get different effects as a pool than you do individually as the clones.

Winzeler: Yes, I suspect you do have some small degree of cross-feeding. You could validate all these responses by testing all your knockout strains one by one, but from what we have been seeing I don't think it's much of a problem, especially with the minimal genes.

Venter: From the transposon insertion mutagenesis we have been doing in *M. genitalium*, we have got down to around 300 essential genes. It bothers me somewhat that this number is similar in yeast. It's certainly a curious coincidence, if not a meaningful one.

Winzeler: There are actually about 1000 essential genes in yeast. I was only discussing the analysis of a fraction of the genome. Even so, there may be fewer essential genes in yeast than in other organisms because of functional redundancy.

Fraser: Did any surprises come out of the subset of genes that appear to be essential? Was there anything you did not expect to see?

Winzeler: It is a big list of genes to go through. We just sporulated the cells under rich medium conditions. In some cases people were able to get knockouts before by adding a specific supplement, such as riboflavin. We observed that some genes were only essential at specific temperatures and in certain strains. In many cases you can figure out why what we found to be essential might not agree with the literature.

Venter: Our two big surprises were that we could knock out tRNA synthases and one specific ribosomal protein that people thought were essential.

Is this list posted anywhere?

Winzeler: Yes, data are available from *http://www-sequence.stanford.edu/group/yeast_deletion_project/deletions3.html*.

Goodfellow: There are several groups that are investing in systems in bacteria where complementation requires multiple compensatory mutations in other genes. This may well be a powerful technology to learn more about physiology.

Venter: These are actual mutations?

Goodfellow: The idea is that you are looking for systems where you can complement with a variety of other mutations, rather than just a single one. The strongest selection is to take out the gene that is essential. The way we normally define 'essentiality' is to score for overnight colony growth. Imagine instead you just leave the bacteria and come back three weeks later and you find there's now something growing. Previously you couldn't easily analyse the bacterium if multiple mutations were involved. But if you have got the whole genome sequence, you just sequence the whole genome again and ask where are the mutations that now allow that thing to grow.

Rubin: This is a common in problem in organisms such as *Drosophila*. If you have a phenotype and keep breeding flies under selection pressure, the phenotype will get less and less severe. If you outcross you get the severe phenotype back.

Goodfellow: That sort of technology might be applicable to try to target large protein complexes, where you know there are 20–30 proteins involved. You can specifically take one out and see if the organism can come up with a solution.

Venter: The biggest dilemma we face in trying to define essential and non-essential genes is determining the pleiotropic partners. This is where

we got into the notion that we had to synthesize a new chromosome with just the 300 essential genes to test the hypothesis, unless you could do the sequential knockouts and have them accumulate. It would be interesting to make a mini yeast, whether with human genes or not, to see whether the cell would actually survive.

Rubin: One problem with making reduced organisms is that genes that you say are non-essential all have some effect on fitness, otherwise they wouldn't be there.

Venter: That is why I said it is context-sensitive: you have to define 'essential' for the situation.

Rubin: In the yeast experiments, suppose you take individual gene knockouts and grow them in these pools for 1000 generations in rich media, what percentage of the knockouts are still there?

Winzeler: We haven't done 1000 generations. We see strong selective pressure. I think 40% of the genes will show some sort of fitness defect for 60 generations. For 1000 generations the number of strains remaining would be small.

Rubin: Of the genes that show a fitness effect after five generations, how many of them will still form a colony on rich media?

Winzeler: They are almost always sick. They grow slowly.

Efcavitch: Can you explain why the deletion phenotype analysis and the expression data didn't quite match up?

Winzeler: You're not necessarily going to see a phenotype with a deletion strain if the gene is redundant. With the gene expression, there are multiple reasons why you might not see the changes at the mRNA level. Transcriptional control is only a small part of gene regulation. This might be more true for yeast than for other organisms, however.

Goodfellow: The *Mycoplasma* experiment showed this: you couldn't see any regulation at all, so there's no correlation between the knockout and the gene expression, because the bacterium has no way of compensating.

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

Patenting genes and gene therapy: legal and ethical aspects

Joseph Straus

Max-Planck-Institute for Foreign and International Patent, Copyright and Competition Law, Marstallplatz 1, D-80539, Munich, Germany

On the difficult, long and costly road from genome to therapy, which involves successful drug development, many factors can prove to be either an obstacle or support. Among those factors patents have from the very advent of recombinant DNA technology occupied a prominent role. In the complex interplay involving research activities in molecular biology, genetics and medicine as well as engineering sciences and applied computer sciences, in the past patents have spurred academic researchers in their research activities and at the same time offered relative security for those private sector investors who were willing to engage and risk money in this new field. Patents have also become a sound pillar for successful cooperation between academic institutions and industry. Because DNA, the molecule that encodes genetic information, is a biochemical substance, it has been treated by patent offices and courts of many countries the same way as any other naturally occurring chemical or biochemical substance. Consequently, between 1981 and 1995 over 1200 patents on human DNA sequences have been issued worldwide and some 5000 respective patent applications filed in the USA alone (Thomas et al 1996, Straus 1997). On the basis of such patents, in the early 1990s successful drugs, such as the production of red blood cell-stimulating erythropoietin (EPO) or granulocyte colony stimulating factor (GCSF) for stimulating the production of leukocytes, reached the market and in the meantime account for far more than US$3 billion of annual sales. Since those patents related to full length genomic or cDNA gene sequences with indicated biological functions, such as encoding various pharmaceutically useful proteins, they were generally viewed as necessary and important, and were met with sympathy not only by

industry, but also by the academic research community (Caskey et al 1995, Deutsche Forschungsgemeinschaft 1997) and remained virtually unnoticed by the general public.

This generally positive attitude towards patents on genes experienced a remarkable change when in 1991 the National Institutes of Health (NIH) filed an application disclosing 3421 so-called expressed sequence tags (ESTs), corresponding to fragments of more than 300 genes expressed in human brain tissue, of which no other function was disclosed than that of being used to identify an expressed gene, or of being used as a sequence-tagged site marker to locate the gene on a physical map of the genome (Caskey et al 1995). Despite the failure of NIH to obtain the patent and its subsequent withdrawal of the application (Straus 1995), the year 1991 clearly constituted a milestone in many respects: the NIH application reflected a new, revolutionary approach to characterizing the human genome through a large-scale (i.e. sequence-based) approach as opposed to the traditional functional approach. Whereas in the latter method, which involves a number of cumbersome steps, at least one function of the given gene is always known, by virtue of the approach this is not true for the first method. This lack of knowledge on biological functions of full-length or partial cDNAs, however, does not automatically deprive the sequences of any commercial value and therefore makes efforts to privatize them understandable (Straus 1996a). Since then, whether or not ESTs should be viewed as patentable subject matter has become a hotly disputed issue of high priority. It brought together an intriguing coalition of pharmaceutical companies, academic researchers and representatives of certain religious groups (Peters 1997), opposing either patenting of genomic and cDNA sequences in general or of ESTs specifically, against a small but potentially powerful group representing the new genomic industry, vigorously claiming the necessity of getting ESTs patented (Straus 1996a, 1997).

Whereas the resistance of religious leaders and other groups of the general public is primarily based on the principle that neither patents on any life forms nor on discoveries should be admitted (Peters 1997), and partly reflects some deficiencies in understanding of the functioning of the patent system (Sagoff 1998), academic researchers as well as the pharmaceutical industry are seriously worried that patents issued on ESTs and single nucleotide polymorphisms (SNPs) could hamper the

development of new therapies (Caskey & Williamson 1996, Marshall 1997, Heller & Eisenberg 1998, Michel 1999). Thus, in their view, ESTs and SNPs, in principle, should constitute and be legally treated as pre-competitive information and should be put into the public domain as soon as possible to maximally stimulate the research that will eventually improve human health (Bentley 1996, Collins et al 1998). To counteract the strong efforts of genomic industries to make this information proprietary, pharmaceutical industry and academic research institutions for the first time have joined forces and recently established a consortium to create a public fine-scale map of the human genome (Wade 1999). Suggestions for treating this pre-competitive information according to models used in the area of copyright, where under certain circumstances rights to remuneration but no rights to exclusivity exist, have been made (Heller & Eisenberg 1998).

The responses of the US and the European law- and policy-makers to the concerns expressed differ considerably as regards the technique applied; whether it will differ in the end result, however, remains to be seen. In the USA the controversial debate eventually ended with no reaction of the law-maker but instead with a decision of the US Patent and Trademark Office, in 1998, to start granting patents for ESTs and SNPs, provided that the usual patentability criteria of novelty, non-obviousness and utility are met. In view of the preceding debates, it has been specifically emphasized that ESTs and SNPs can meet the utility requirement also by specific utilities, such as to be used to trace ancestry or parentage (SNPs) or for chromosome identification and gene mapping (ESTs) (Doll 1998). On the other hand, the European Union in its Directive on the legal protection of biotechnological inventions of July 6, 1998 (98/44/EC) (European Union 1998) paid specific attention to the concerns expressed and introduced a whole set of rules aimed at balancing the interests at hand: first, it clarified that the simple discovery of the sequence or partial sequence of a gene cannot constitute a patentable invention. Without indication of a function a mere DNA sequence does not contain any technical information and thus cannot be viewed as patentable invention. However, if isolated from the human body or otherwise technically produced this may be the case, even if the structure is identical to that of a natural element. All that, provided that industrial application is disclosed in the patent application. In cases where

a gene (also partial) sequence is used to produce a protein, it has to be specified which protein (or partial protein) is produced or what function is performed. Since under the 'indication of a function' not the indication of a biological but of any function responsible (causal) for a technically applicable result has to be understood, ESTs and SNPs, which can be used, for instance as diagnostic markers or for identification for forensic purposes, in principle are eligible for patent protection in Europe, too (Bostyn 1999, Oser 1999, Straus 1998). Moreover, the European law-maker also made an attempt to solve or at least ease the problem of dependency, which in respect to DNA patents has to be viewed as most serious: when patented sequences overlap only in parts which are not essential to the invention, each sequence is to be considered as an independent sequence in patent law terms. The success of this rule will to a large extent depend on how the term 'essential' will be interpreted (Straus 1998).

Once the thorny road from genome to gene therapy is mastered (i.e. the gene or genes of interest discovered and sequenced, their function cleared up, and transfer and expression vectors successfully constructed and tested), researchers again will be faced with seemingly widely differing legal situations in the USA and in Europe. Whereas in the USA inventors will not experience any specific limitations and might well be issued process or method claims of remarkable breadth, covering practically any imaginable *ex vivo* somatic gene therapy, as was the case with the US patent No. 5.399.346 of Anderson et al (Flanagan 1998, Straus 1996b), in Europe they will have to learn that therapeutic methods are not regarded to be susceptible to industrial application and therefore not patentable. This, however, does not mean that they will be left without any protection. Since substances or compositions for use in such methods are eligible for patent protection, not only methods for their production but also intermediaries and, eventually the endproduct—the drug itself—involved in somatic gene therapy and somatic cell therapy, such as vectors, somatic cells, as well as transformed somatic cells, to be injected, infused, etc. can be patented. Outside patent protection, thus, remain only entire therapeutic methods, including the steps of removing human tissue and injecting, etc., the drug (Bostyn 1999, Straus 1996b). Neither in Europe, nor in the USA, however, will patients treated with and physicians applying

gene therapy be affected by patents related to gene therapy, at least as far as the *ex vivo* therapy is at hand. In Europe, the rights conferred by a patent do not extend to the production of drugs for direct use applied to individual patients on medical prescription (Bostyn 1999) and in the USA changes introduced into the Patents Act in 1996 deprived patentees of remedies for infringement by a medical practitioner's performance of a 'medical activity' (35 USC 287). Moreover, the EU Biotech Directive explicitly exempted the human body, at various stages of its formation and development, from any effect of a patent.

At this point in time it seems premature to predict whether the European legislative approach or the more pragmatic and flexible US approach will provide for the necessary degree of protection yet also give adequate incentives to all actively paving the way from genome to therapy. Should the patent system fulfil its primary goal, namely to foster innovation to the benefit of society as a whole, those incentives will have to be commensurate to the respective contribution to the art and should not be available for speculative 'achievements'. Efforts, such as the SNP consortium, which will rapidly enlarge the state of the art, will certainly reduce the prospects that minor achievements could seriously impede drug development. On the other hand, the beneficial effects of the patent system with respect to competitive information will not be affected and will remain fully operational.

References

Bentley DR 1996 Genomic sequence information should be released immediately and freely in the public domain. Science 274:533–534

Bostyn SJR 1999 The patentability of genetic information carriers. Intellectual Property Quarterly 1:1–36

Caskey CT, Williamson AR 1996 Merck, SmithKline and patents. Nature 381:360

Caskey CT, Eisenberg RS, Lander ES, Straus J 1995 HUGO statement on patenting of DNA sequences. Genome Digest 2:2

Collins FS, Patrinos A, Jordan E, Chakravarti A, Gesteland R, Walters L 1998 New goals for the US Human Genome Project: 1998–2003. Science 282:682–689

Deutsche Forschungsgemeinschaft 1997 Genforschung—Therapie, Technik, Patentierung. VCH, Weinheim (Senatkommission fur Grundsatzfrgen der Genforschung)

Doll JJ 1998 The patenting of DNA. Science 280:689–690

European Union 1998 Directive 98/44/EC of the European Parliament and of the Council of 6 July 1998 on the legal protection of biotechnological inventions, OJ EC No. L 213/13 of 30/7/98. European Parliament/Council of European Union

Flanagan JK 1998 Gene therapy and patents. J Patent Trademark Office Soc 10:739–751

Heller MA, Eisenberg RS 1998 Can patents deter innovation? The anticommons in biomedical research. Science 280:698–701
Marshall E 1997 Whose DNA is it, anyway? Science 278:564–567 (erratum: 1997 Science 278:1551)
Michel H 1999 Die Sequenzierung des Humangenoms—Auf's falsche Pferd gesetzt? Biospectrum 1
Oser A 1999 Patenting (partial) gene sequences taking particular account of the EST issue. Int Rev Ind Property Copyright Law 30:1–18
Peters T 1997 Playing God? Genetic determination and human freedom. Routledge, New York
Sagoff M 1998 Patented genes: an ethical appraisal. Issues Sci Technol 14(3): 37–41
Straus J 1995 Patenting human genes in Europe—past developments and prospects for the future. Int Rev Ind Property Copyright Law 26:920–950
Straus J 1996a Intellectual property in human genome research results. OECD STI Review 19:45–64
Straus J 1996b Patentrechtliche Probleme der Gentherapie. GRUR 10–16
Straus J 1997 Genpatente: rechtliche, ethische, wissenschafts—und entwicklungspolitische Fragen. Helbing & Lichtenhahn, Basel
Straus J 1998 Abhängigkeit bei Patenten auf genetische Information—Ein Sonderfall? GRUR 314–320
Thomas SM, Davies AR, Birtwistle SM, Crauther SM, Burke JF 1996 Ownership of the human genome. Nature 380:387–388
Wade N 1999 10 drug makers join to find diseases' genetic routes. The New York Times 4/15/99 pg 27

DISCUSSION

Venter: In the USA, one of the key issues is the dependency issue you raised. I recently had a meeting with the US patent commissioner on this issue, and it looks as if the US Patent Office is going to go the conservative route, with comprising versus containing language. People who have studied the patents issued to Incyte view them as having little value, because they were just on the limited sequences: they don't get the whole gene from those. For example, many of the sequences that they claim were for kinases were not, and the sequences were of relatively low quality so they got the composition of sequences that probably don't exist and that's what their patents cover. I think the US Patent Office will be taking a very limited view. Therefore, anybody who gets the full length sequence won't necessarily have a dependency, but it still hasn't been determined.

Souza: Most people have taken a wait and see approach.

Venter: The ruling is going to come out soon—at the very least, the guidelines will be issued soon. There has been a slow-down on DNA patents coming out of the Patent Office recently while they're trying to decide these issues. I don't think there's any question that ESTs and SNPs

and other parts of DNA will be patentable, it is just that the claims are going to be limited. One of the things that I don't think is widely understood about the SNP consortium, is that they are actually going to be filing provisional patents on what they discover, with the claimed intent to turn those into statutory invention registrations. It is not clear whether this is going to be possible. It is not as though they are just going into the public domain. There are other implications to this.

Straus: But they wanted to do that only in order to prevent others from patenting, I think.

Rubin: Can you say what you mean when you say that the claims are going to have a limited scope?

Venter: The NIH claims that it was fairly obvious to people that you could use the EST to get the complete gene as a research tool. NIH tried to claim the entire gene from the EST and Incyte and other organizations tried the same thing. But by a conservative approach I guess that the Patent Office are taking the view that it is a method to get the entire gene, but it doesn't tell you what the composition and the matter of that gene is, and therefore it is not predictive and the EST does not give you the entire gene as a claim. Whatever that EST is worth as a tool, as a marker or diagnostic test, for example, is the limit of what the claims are. Thus my understanding is that it doesn't create a dependency if somebody else clones the entire gene by a different method than using the EST as a probe.

Rubin: But would it prevent someone from making an expression microarray using the EST sequence?

Venter: That is one of the areas that is absolutely uncertain, according to my understanding.

Straus: We don't have a court decision for that, and as long as we don't have such decisions, one can only speculate. The entire problem is that you can get patents also for further development, but those patents will possibly be dependent on the first one. Only in the case when infringement is at stake, will the courts decide the scope of protection.

Venter: With SNP patents, nobody even knows what that means. The SNP consortium are going to take 50 base pairs on either side of the variation sites, and file a provisional application for a 100 base pair sequence. Celera indicated that it was going to take the same approach, when we talked to the patent officer about this. But for tens of millions of

SNPs, what does that mean? We're actually pushing for the same sort of limited value to SNP patents as with ESTs. If somebody comes up and says they have the data to show that this particular SNP actually determines whether or not somebody responds or doesn't respond to a certain drug, that's a new discovery or invention on top of things, so it doesn't matter whether it's in the public domain or in our patent set. It doesn't discourage further invention. The question then becomes one of dependency issues.

Rubin: What about research use, using your SNP in an experiment to determine whether it's linked to something or not.

Straus: I would say that is being used as a tool, and is not covered by the classical understanding of research exemptions. It is like using PCR. Therefore my question is, do we need relevant statutory provision on the research exemption, which would cover not only acts done for experimental purposes relating to subject matter of the patented invention, as it already exists in the European Law, but one which would also cover the use of the patented invention as research tool? The German Research Foundation (DFG) has asked for that. The problem is how to differentiate between researchers working in academic areas and industry. Would that be justified?

Goodfellow: The practice is more complicated, anyway. For example, it is common practice to make a compound for testing whether it has the properties which are claimed for it in the patent.

Straus: That is covered in Europe by the research exemption rule, but not in the USA.

Venter: We're actually trying a new approach with the provisional patents, to make the provisional filing date available to any of the Celera database subscribers, just to use to build their own intellectual property, not as a weapon against them but as an advantage for them.

Straus: You must not forget (and this is a real worry) that the subsequent filings based on the provisional applications will be limited to the disclosure contained in that provisional application. You cannot go beyond that later on, or only with a later priority date. New matter added to what has been disclosed in the provisional application may turn out not to be patentable, if you don't have a grace period like there is in the USA. Then the added matter may be 'obvious', in view of what has been disclosed not only in the first patent application but also

otherwise published. This is because the publication will form the state of the art for all later filings. This is the situation in Europe.

Venter: That's one of the key differences between patents in the USA and Europe.

Magnus: In the USA in case law there is something called the 'product of nature' doctrine. Do you have something analogous to this in the European statutory system that is not actually statutory?

Straus: No, but as you have seen, we have just the opposite: we have a statutory provision that you can patent something which is identical to something which existed in nature before. Our case law has developed in a different direction than that of the USA, and has been approved by the EU Biotech Directive.

Magnus: In addition to the controversies over ESTs and SNPs there have been controversies over disease gene patents. One of the attacks on these is that they violate the product of nature doctrine. In Europe do you have a similar debate?

Straus: Amgen had a problem with the product of nature doctrine. Probably to avoid the product of nature objection, the erythropoietin (EPO) patents were originally filed in a way to show that there are substantial differences between the EPO produced by recombinant DNA technology and the naturally occurring EPO. This caused a great deal of difficulty when they applied for FDA approval, since they had to show that their EPO has exactly the same properties as that occurring in nature! Patent lawyers spent some time in explaining those data as measurement errors.

Souza: I would like to ask a question related to antibodies. In the 1980s, patent claims to antibodies were limited to the specific monoclonals. In the late 1990s granted claims have given protection to all monoclonals that are directed to a specific antigen: the Patent Office has done a 180° turnaround. Will these new broad claims stand up in court, or are we likely to see a narrowing of claim structure in the future?

Straus: I can only comment that the European Patent Office is trying to narrow the claims to what has been disclosed. It may be that the European Patent Office is more liberal now in that particular area, and then later on if the disclosure is not viewed as enabling across the whole breadth of the claim you can lose your patent.

Venter: But it's disturbing, because a lot of people use the arguments with monoclonal and other antibodies, saying that you don't have to worry about DNA patents, because look what happened to monoclonal antibodies—they basically became non-patentable for a while. Now, as Larry Souza has pointed out, the situation is reversing itself again. This doesn't give a lot of comfort to people who think that DNA patents will largely go away as an issue.

Straus: They will not go away for the time being or for the foreseeable future.

Magnus: Aside from the legal issues of patents, which are important and on peoples' minds, a separate issue is how we want to behave. Just because the courts say that certain materials are patentable doesn't mean that these are things that people ought to patent as a matter of practice. For example, the AMA code of ethics prohibits that physicians should patent any kinds of procedures.

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

Ethical issues: from genome to therapy

David Magnus

University of Pennsylvania Center for Bioethics, 3401 Market Street #320, Philadelphia, PA 19104-3308, USA

In considering the ethical issues that the development of genetic technology presents, it is useful to divide their impact on medical practice into relatively short term (the next five years) and relatively long term (up to 20 years and beyond). Both sets of issues need immediate consideration and action and there are important connections between them.

The most pressing short-term problems arise out of the fact that the most immediate impact of finished sequence data on the human genome is likely to be in diagnostics, with the primary therapies being prenatal diagnosis and abortion. There are a large number of important issues that have been raised with regard to genetic testing. These include worries about the quality of informed consent when dealing with complicated statistical and probabilistic information and limited access to genetic counselling; worries about genetic privacy and confidentiality and genetic discrimination; and concerns about a new eugenics as testing slides from cystic fibrosis to baldness, a tendency toward obesity, homosexuality, and other value-laden traits (for example, Andrews et al 1994, Kevles & Hood 1992, McGee 1997).

These worries are important and a great deal has been written about them (Magnus & Butcher 1999). I wish to develop some discussion of the reflection of these concerns as they relate to more long-term technological development. We will see that many of the current problems will be greatly exacerbated by the increasing power we will develop to understand and control genetic material.

I am optimistic about the long-term prospects for very powerful therapies and the potential for profound genetic manipulations that will produce an explosion of pharmaceutical developments and new kinds of therapies. Even more radically, the genomics of the future (10–20 years from now) will no longer be concerned primarily with sequencing but

with creating sequences, base pair by base pair. This will allow the creation of artificial chromosomes 'from scratch' and combined with developments in nuclear transfer will make a radical transformation of gene therapy possible. The potential to design synthesized life forms similarly will beckon a new era in biotechnology.

However, this rosy vision has a price. There are a number of issues that these developments will raise—issues already visible with current technology. There are going to be concerns over ownership and control of living things and of genetic material. The patenting of living organisms and of genes is widely practiced, well established, and still highly controversial with a potential for public backlash.

Another significant concern that is often neglected is that the reductionist tendencies of both our culture and our science may mislead us about how much power we really have and to ignore very powerful causal factors which do not fit within our reductive models. For example, we give far too much credit to the development of vaccines and antibiotics for the decline in mortality that has occurred in the past century and to our apparent triumph over the most common infectious diseases. In fact, only a very small part (between one and five per cent) of the decline in mortality can really be attributed to discrete medical interventions. It is important that our enthusiasm for genetic technology does not lead us to overestimate (or oversell to the public) how much we can do, nor do we want to underestimate the importance of environmental and social conditions on the health of populations.

Another worry that will become increasingly important is the 'backlash' problem which biotechnology faces. The public is increasingly responding to the rapidly expanding power of biotechnology in a visceral way—raising concerns over 'playing God', 'acting unnaturally', etc. This public response can potentially be dangerous for many prominent technologies and therapies. It has already created problems for genetically modified crops in Europe and has the potential to create regulations that are counterproductive.

Typically, there are real issues behind the often overblown and misguided alarm raised by the public. In particular, three concerns that exist now are likely to be exacerbated by the genomics of the future. First, the fear of a new eugenics will become much more pronounced. Increasing knowledge of genetic contributions to many more traits

combined with the possibility of designer chromosomes will make it possible for at least some people to choose the traits of their offspring. The potential for the creation of a genetic underclass and the stigmatization of much of the variation that makes life interesting is a frightening prospect to most people.

Second, an increasing number of people worry about the potential ecological harm which may be wrought by the creation and introduction of new life forms into the environment. Recent studies that revealed the potential harmful effect of some common forms of genetically modified corn on Monarch butterflies highlight the need for careful thought about creation and use of new life forms. This type of ecological concern is particularly acute in Europe, but there are signs that it is growing in the USA. The vast power of the new genomics is likely to make this a major issue everywhere.

A third issue that the new genomics will raise is the increased potential for biological weapons—potentially even weapons of mass destruction. When it is possible to synthesize novel genomes, it may be possible to design them to bring about death, either indiscriminately or targeted at specific groups.

Although each of these issues deserves serious attention, it would be a mistake to curtail promising scientific research with enormous potential benefits simply to forestall the potential pitfalls. At the same time, the backlash problem makes it more likely that regulations may be adopted which throw the baby out with the bathwater. What is called for is a set of solutions that can be helpful in avoiding the backlash problem, dealing responsibly with the very real dangers of the technology, and still maintain the development of the technology and the science behind it.

There are three things that can be done to help address these issues. First, it is imperative that efforts be made to better educate the public about genetics, and the ethical issues that it raises. The more informed the public, the more realistic the assessment of the costs and benefits of technology, the less likely that there will be the kind of visceral backlash that so often occurs.

Second, professional groups of scientists and clinicians need to do a much better job of self-regulation. When the public (at least in the USA) became anxious over recombinant DNA technology, the genetics community agreed upon a temporary, self-imposed moratorium on

rDNA research. Subsequent research proved that the technology was actually quite safe. Many (for example, James Watson) have claimed that the lesson to be learned from this story is 'never again'. They claim that scientific progress was held back by this action for what turned out to be unfounded alarmism on the part of the public. I would argue that this view is misguided. Without the moratorium, regulations might well have been legislated that would have been far more cumbersome for science and much more difficult to eliminate. Self-regulation is important and it is not widely practiced.

Third, in a similar vein, science and industry both need to do much more prophylactic bioethics. The cloning of Dolly was a good example of what should be avoided—research likely to elicit strong public response, without adequate thought to the ethical issues in advance. More scientists and industries need to allow for bioethical reflection prior to the announcement of new developments to an increasingly concerned public. This allows for a much higher level of discourse and is much more likely to avoid the backlash problem.

References

Andrews LB, Fullarton JE, Holtzman NA, Motulaky AG (eds) 1994 Assessing genetic risks: implications for health and social policy. National Academy Press, Washington, DC

Kevles D, Hood L (eds) 1992 The code of codes: scientific and social issues in the Human Genome Project. Harvard University Press, Cambridge, MA

Magnus D, Butcher A 1999 Contemporary genetic technology: scientific, ethical and social challenges. Krieger Press, Florida

McGee G 1997 The perfect baby: a pragmatic approach to genetics. Rowman and Littlefield, New York

DISCUSSION

Goodfellow: I'm not totally convinced about the ideas of self-regulation. Although all of us subscribe to the view that everyone has to have a moral standard for themselves, I am not convinced that the British Medical Association, for example, is the right body to be imposing the ethical standards for doctors—perhaps patients should have a view. What often happens with self-regulation is that it occurs behind closed doors, and you don't get the public debate, which in the end you can never avoid.

Magnus: At a fundamental level, I disagree with you, in that it seems to me that the professional associations and societies are the people who know best the practices and activities of each profession.

Goodfellow: They can also be the most conservative elements in society.

Venter: In a retrospective fashion, yes. I would take Peter Goodfellow's issue even further. You recommended that scientific journals should be involved in regulation: for example, if Gerry Rubin patents a *Drosophila* gene, they shouldn't publish the paper. But journals are commercial entities: it makes no sense for them to be the ones to establish ethical standards. After all, they have trouble enough setting scientific standards with peer review.

This is what we are seeing right now with the genome centre at NIH. Because a few people there think that patenting DNA is fundamentally wrong, they're making every attempt to thwart it, even though the only purpose for patenting human DNA is to create new diagnostics and therapeutics. One person's ethics is another person's way to change society. To leave it up to scientific societies or journals is highly questionable.

Magnus: This is why we have to look to areas where consensus is reached.

Venter: One way in which bioethics can serve the community is in the area of gene therapy, working through the issues raised by enhancement gene therapy, distinguishing between things that are totally hypothetical versus things that are real. I am not sure there are too many instances where patents have truly inhibited basic research or blocked therapies from going forward.

Magnus: That is now happening. We had to discontinue *BRCA* gene testing at the Hospital of the University of Pennsylvania because of a lawsuit. It's not very hard to imagine restrictions on research in the area of genetic testing.

Venter: It is not blocking people from doing the test, it is just that your hospital is not able to do it, charge for it and make money on it.

Roses: Every action has an equal and opposite reaction. One thing that has happened with the huge amount of ELSI funding from the NIH Genome Project going into the 'ethics industry', is that we now have another interest group. The interest group, in many cases, does not speak for the public by using scholarship to gauge what the public's view is in terms of data and evidence, but as groups of ethicists they decide what the public thinks. When we talk about a backlash, what we're talking about is a journalistic reaction. In terms of the great unwashed public, most of them are not properly educated or accurately

polled. We look at polls that ask questions in such a way that they influence the responses. The kinds of feelings that are actually in the population really need to be studied in a quasi-epidemiological way, so people can actually speak on behalf of the public.

Magnus: I have two responses to that. First, bioethicists do empirical research. There is a lot of study of public, patient and physician attitudes. Second, I think you are mistaken if you think that backlash is simply a media creation. Their job is to sell newspapers. If the public really didn't respond to these issues, they wouldn't sell them. Figuring out cause and effect here is very difficult, but in the man-on-the-street interviews you see on TV shows every time these issues are raised, overwhelmingly you see the same kind of responses.

Venter: This was put to a real test right here in Switzerland in a public referendum.

Hochstrasser: This was very scary. I realized Switzerland was conservative, but in certain areas more than 60% of the people voted against any molecular biology! Eventually, the initiative was banned, but it was cause for real concern.

Venter: All of science needs to do a much better public education job. However, a lot of these concerns are creations of the press. My favourite is one that Art Caplan and I wrote an editorial on, entitled 'Using one's head' (Caplan & Venter 1997). It came out of a British laboratory where a journalist was asking a researcher working on salamanders what possible applications his work would have for medicine. He said that they could perhaps apply the same technology and clone headless humans for spare body parts! This made national evening news in the USA. It was on CBS evening news, and was made even worse by a former NIH director doing a major public commentary saying, 'The worse problem, Dan, is that these people wouldn't have the opportunity to say no'!

Roses: When a thing like this happens, the public is educated through the journalists.

Venter: Art Caplan and I pointed out in our editorial that most students of biology and medicine have a fundamental understanding that humans without heads are dead. There is fundamental biology missing in these scare stories. I was riding in a taxi shortly after this was on the news (that's how I take my poll of what the public thinks), and the driver was horrified: 'Did you hear about this? These guys are going to start cloning

headless humans!' Something like this really catches on. The total solution is not to cut off the press, but I think it means we have to do a better job of educating the public.

Mann: With regard to public education, we would like to believe that things get better if we educate the public, but I am aware of studies in Denmark that have shown that education actually doesn't make a difference in the rejection rates.

Herrling: In Switzerland, in our recent referendum, the people who knew nothing about gene technology were more positive, in general. A little bit of information may be more dangerous than no knowledge.

Mann: In Germany about 10 years ago there was a case where a scientist in Heidelberg went out of his way to explain his controversial work to the public only for him to have his lab blown to pieces.

Magnus: I think this shows that education is more than just having some information. Following up some of the discussions of the media, I think we also need to hold journalists more accountable. We have just launched an empirical study of the Kevorkian story on *60 minutes*. Of the 1400 stories that appeared in the media overall, that story was covered overwhelmingly as a crime story. Less than 3% of stories that came out dealt with any of the relevant ethical issues or arguments. So clearly, more needs to be done. I don't think it is because the media is vicious or malicious: they are clearly sensationalistic, but it is important not to underestimate just how ignorant they are. They know nothing about science or medicine. Many participants here have done a lot of media work. How many times have you talked to producers who don't understand the first thing? I have had reporters literally ask what DNA is, and these are people writing for major newspapers. We need to hold them accountable when they don't do their job, and educate them so they can do a better job.

Straus: Just a word about self-regulation versus the statutory approach. The statutory approach has the disadvantage that it becomes very inflexible. For example, in Germany we adopted the Embryo Protection Law before anybody knew anything about potential uses of embryonic stem cells. Now, we may have a serious problem if it turns out that the law does not offer adequate solutions. What makes me somewhat nervous is listening to observations on the negative impact of patents on research and exchange of scientific information and how people should behave.

People will always behave like humans do. If you have something which is close to commercial exploitation, you will have a lot of people who will go for patents, no matter what people think. And therefore my question is, why is nobody asking in the USA for a statutory rule on research exemption? In Europe, as I indicated, we have explicit statutory rules and we have case law in the UK, Germany and other countries that you can use patented products in order to test them, or to improve them, no matter what the eventual goal of your activity is. Every time I raise this question in the USA, even the researchers remain silent.

Venter: There's a good reason for that. Other than potentially with PCR, it's a non-issue. This is one of those hypothetical issues that everybody runs around wringing their hands over, and there have been a few cases where researchers were threatened to be sued for using a technology as a means of controlling things. For Roche and Perkin Elmer, PCR is one of their biggest products. Basically, this is a case where one person's problem, their desire to make a reagent in their lab, is another person's industry. We can make Taq polymerase in a short time in the laboratory, but the PCR patents are valuable in terms of diagnostics and other areas, and we've been told we can't make our Taq polymerase in the lab for any purposes. But let's not forget that we benefit from the commercialization of reagents. We can buy any restriction enzyme, for example, from commercial companies instead of having to spend most of our time in the lab making reagents.

Straus: Here we are talking about the use as a research tool. However, you are not doing any work in further developing the PCR technology, or Taq polymerase. This kind of use is not covered by the statutory research exemption rule in Europe either.

Rubin: As an academic researcher, I don't mind the patents on Taq polymerase because I can buy it and Roche doesn't ask for any rights to anything I invent using it. I may pay 100 times more for the enzyme that I might if there were no patent, but I don't have any restrictions on the use. It would be very different if Roche said they want 1% on each of the royalties on anything you discover using PCR. This would be much more problematic. This is the kind of patent that I worry about.

Venter: This is not really about patents. It is the licensing strategy that the person with the rights to the patent pursues. This was the case with Human Genome Sciences. They required extreme reach-through rights

of somebody using their database. It does not have to work this way. Our model for Celera does not have reach-through. When we talked about the formation of Celera with the director of the NIH, Dr Varmus, he said this is fine. He saw it like the situation of the commercial production of restriction enzymes. As in the case where the restriction enzyme seller doesn't own a piece of your experiment that you use the restriction enzymes for he was very supportive. This is how we are setting up the database: it is free for people to use for inventions—there is no reach-through. These are not patent policies. This is about business practices and business policies.

Magnus: But they are made possible by the fact that when you have an exclusive right to genes and their uses and products, you can tie up all uses and thereby tie up future useful discoveries. With the Myriad patents, people who have control and exclusive rights to certain genetic tests basically said we will allow clinical genetic testing, possibly even for research purposes, but any results that might accrue belong to Myriad.

Venter: Other than for *BRCA1*, nobody seems to worry about this. SmithKline Beecham has a huge diagnostic enterprise with patents on all the diagnostics, as does Abbott. It is only an emotional issue because it is the *BRCA1* test and one company has intellectual property on that. It's not blocking anybody from doing *BRCA1* tests for research purposes.

Goodfellow: I can understand the frustration, but I don't see the difference between the *BRCA1* test and buying a drug.

Magnus: There are several differences. First, public money raises funds that go towards the creation of other drugs and one of the justifications in the Bayh–Dole act in 1980 was that Congress wanted to fill the gap between basic science and the practical benefits to be created. When we are talking about disease gene patents the gap between the basic discovery and the product is much shorter. When you have the basic sequence information it is relatively trivial to get to the test.

Lipshutz: It is not a trivial transition to get something from the published article into a product that has the reliability needed as a clinical test.

Reference

Caplan A, Venter C 1997 Using one's head. Science 278:1547–1548

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

The impact of genomics on drug discovery

Peter N. Goodfellow

SmithKline Beecham Pharmaceuticals, New Frontiers Science Park, Third Avenue, Harlow, Essex CM19 5AW, UK

In the 1980s, DNA cloning methods allowed the facile production of protein targets *in vitro*. This technology also resulted in a modest increase in the targets available for screening. Nevertheless, the number of different targets studied by the pharmaceutical industry in the 50 years from 1940–1990 was limited to only about 500. Genomics has alleviated the problem of target limitation. The Human Genome Project and cDNA sequences available from a variety of sources (Adams et al 1995) have provided access to over 50 000 human genes. The sequencing of bacterial genomes has caused a similar explosion in available targets for antibiotic research.

Making a new pharmaceutical product is expensive and takes a long time. The total spent on research and development by the major pharmaceutical companies in 1997 has been estimated to be US$35 billion (Kettler 1999). In the same year, 38 new chemical entities were approved for sale. This suggests that the cost of bringing a new drug to market is approximately $1 billion. More sophisticated estimates have proclaimed figures of about $600 million per drug, and the costs are increasing over time (Kettler 1999). The drugs on sale today were discovered between one and three decades ago; the average time to bring a new drug to market is probably more than 12 years. These observations pose three obvious questions:

(1) Why is the process so expensive?
(2) Why does it take such a long time?
(3) Why is the cost of producing new drugs apparently increasing when implementation of new technologies should lead to a decrease in cost?

The major drivers of the increase in both time and cost are the size and complexity of development programmes. Traditionally, the development phase has consumed two-thirds of the research and development budgets. Similarly, the development phase is usually twice as long as the research phase. Any new technology that improves cycle time and cost in the discovery phase will have a smaller overall impact than technologies that affect the development phase.

Drug discovery and development requires the integration of a large number of techniques from a variety of biological, chemical, physical and medical disciplines. To be successful, a company must either achieve expertise in all necessary disciplines or resource the same expertise from outside. At any point in time, new technical advances offer hope that the efficiency of drug discovery and development can be improved. However, before these hopes can be realized, new technology must be integrated successfully into the process. The costs of producing new drugs are increasing because the costs of development are increasing, and because of the cost of maintaining the complex set of skills needed in platform technologies.

Genomics, by increasing the spectrum of available targets for drug discovery, will improve the chances of producing novel pharmaceutical products. However, a genomics approach is predicated upon improvements in bioinformatics and target validation techniques.

References

Adams MD, Kerlavage AR, Fleischmann RD et al 1995 Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature (suppl) 377:3–174

Kettler H 1999 Updating the cost of a new chemical entity. Office of Health Economics, London

DISCUSSION

Mann: When I talk to the people in the trenches at pharmaceutical companies, they sometimes give a different impression, that not so much has changed, just the expectations. Combinatorial chemistry is a case in point; the claim was that we could deal with millions of molecules in parallel. I have heard even from companies that are specializing in this approach that they have now given up on it. Aren't

we now back to where we started, with one chemical entity synthesized at a time?

Goodfellow: I don't think that's true. Everybody is struggling to integrate new technologies all the time, and certainly the output of compounds per chemist has increased dramatically in the last three or four years. The technology is there, but putting the technology together is a challenge to all of us. A pipeline will only run as fast as the biggest constriction. My own view is that the limiting factors for making new drugs are not the biology, but the chemistry.

Roses: The business of targets really is a matter of focus, now. With 100 000 targets, what can you target, and what is valuable? All targets are not created equal. It costs a tremendous amount of money to invest in a target.

Fraser: What's your best guess of the number of viable drug targets? Is it going to increase twofold or fivefold, for instance, from genomics?

Goodfellow: Everything which was possible to study was previously used as a drug target. The number of 400 I gave covered all the available proteins in 1990.

In the same way that Craig Venter and you have dreamed about the world, I have dreamed about a world where we start with a drug for every protein. There is a solution to the problem that says, 'Let's make a drug for every protein, and then test the drugs'. I suspect that every protein, if you modify it, will change the physiology of the organism. And, if that's true, you may find a condition where that modification can bring you benefit.

Venter: When we look at changes that are going to take place in society, it will have a big impact when people begin to realize that only roughly one-third to one-half of a population given a drug has a good response. This will be one of the biggest changes in terms of how the SNP-based individualized medicine paradigm fits in.

Goodfellow: I'm not sure. You reach a certain time in life when it's time for young people to do the experiments, and I've reached that time. I grew up with human genetics when we were trying to combine it with molecular biology. Conventional wisdom stated it wasn't possible to map genes to human chromosomes, it was not possible to do linkage studies in humans and human disease genes could not be cloned using only information about chromosomal location. Conventional wisdom was overturned. Now I

have to confess that I may be suffering from a bad case of conventional wisdom: I'm not sure that we're going to be able analyse, using genetic techniques, all common genetic diseases. For a few diseases, it will be possible, but I'm not sure anymore whether we will have the power for most diseases. I base this negative view on experience with multiple sclerosis (MS). Almost every family in the UK where there was more than one affected sibling was collected, and used in a linkage study. The study showed that MS was a complex genetic disease, but it was not possible to identify any specific gene which was involved in the disease. A recent meta analysis of families from both Europe and the USA still did not have enough power to identify specific loci. Whether associations studies will provide the power needed, we don't know. It may prove impossible to analyse complex phenotypes in an outbred population using these approaches. This is why young people should do these experiments.

Venter: I guess I'm still young at heart, because I believe that as the number of markers increases, the power increases and also our ability to resolve things increases.

Roses: When you go to see a doctor and there's something medically wrong, you want accurate diagnosis and 100% effective therapy with 0% risk, and you want it for free. What as a society are we willing to pay to take the danger out of the game, and what are we willing to do to make it economically more feasible to test the number of targets that are available? These are the kinds of issues that we need to deal with. And as I said yesterday, when it comes to the mass marketing of abstracted SNP profiles, they really give no genetic information other than the response to that particular molecule.

Goodfellow: Essentially the amount of money which is spent on drug discovery has fallen. It used to be the rule of thumb that the ratio was 2:1 development to discovery. It is now running more like 3 or 4:1. Allen Roses is right; if you don't get a short-term benefit in terms of producing drugs, probably the biggest benefit will come from doing something about the cost of clinical trials.

Hochstrasser: An intellectual property question: if a gene is known and patented, and someone discovers a protein from the proteomic side, what happens with the intellectual property?

Straus: If the protein is patented as an expression product of the gene, the protein is covered no matter how produced. If only a part of the DNA

sequence or even the full-length DNA sequence is patented, for instance as a marker, but the protein at hand was neither claimed nor disclosed, then the patent should not cover the protein. Depending on the relevant state of the art, the protein could be patented and the patent should not be dependent on the gene patent.

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

From genome to therapy: industry perspective

Paul L. Herrling

Novartis Pharma AG, Research, CH-4002 Basel, Switzerland

Genomics and its impact on pharmaceutical discovery efforts

The Human Genome Project is an international endeavour aiming at detailed genetic and physical maps of the human genome and sequencing of the DNA for the estimated 100 000 genes it contains (Watson 1990, Cohen et al 1993, Fields et al 1994). Genomics research has demonstrated that many diseases, such as Alzheimer's disease (Yankner 1996, Selkoe 1997), Parkinson's disease (Goedert 1997, Polymeropoulos et al 1997), diabetes (Kahn et al 1996), asthma (Postma et al 1995) and rheumatoid arthritis (Maddison et al 1993) have an important genetic component which interacts with environmental factors to manifest the disease state. In many instances it will be more straightforward to address the underlying genetic basis of the disease rather than to change the environment and human behaviour.

If one considers the development of biomedical research since the 1970s, when relatively few well-characterized enzymes and receptors were available to pharmaceutical researchers, advances in molecular biology and recombinant DNA technology in the 1980s resulted in an exponential growth in the number of genes and gene products available as possible disease targets. Functional genomics can be defined as the scientific activities that are being applied to link genomics research with the process of discovery of disease-relevant therapeutic targets. Since not even the largest companies have resources to address all potential disease targets, those that can rapidly identify and assign function for key disease genes and proteins, and accordingly focus their activities, will achieve a lead over their competitors.

Building a competitive functional genomics platform

Currently, marketed drugs interact with *c.* 400 genes or gene products, and estimated numbers of important genes for disease predisposition, onset and progression range from 3000–10 000. Therefore, many novel genes and proteins have yet to be identified as proprietary targets for pharmaceutical research.

Novartis' decision to build up strong in-house functional genomics (Fig. 1) took into account several strategic factors (Dyer et al 1999a,b).

- Since success of functional genomics in the pharmaceutical arena will be measured in terms of productivity in contributing validated, disease-relevant targets which can then be exploited for therapeutic discovery by Novartis' disease-oriented research groups, strong in-house groups are necessary to achieve this goal (Fig. 2).
- Within Novartis Research, several programmes existed with established expertise that could be re-focused into the new genomics group. This allows critical mass allocation to key problems in the field, such as high-throughput functional profiling methodology for gene function.
- Although functional genomics approaches are the focus of dedicated efforts in the biotechnology sector, the majority of these smaller

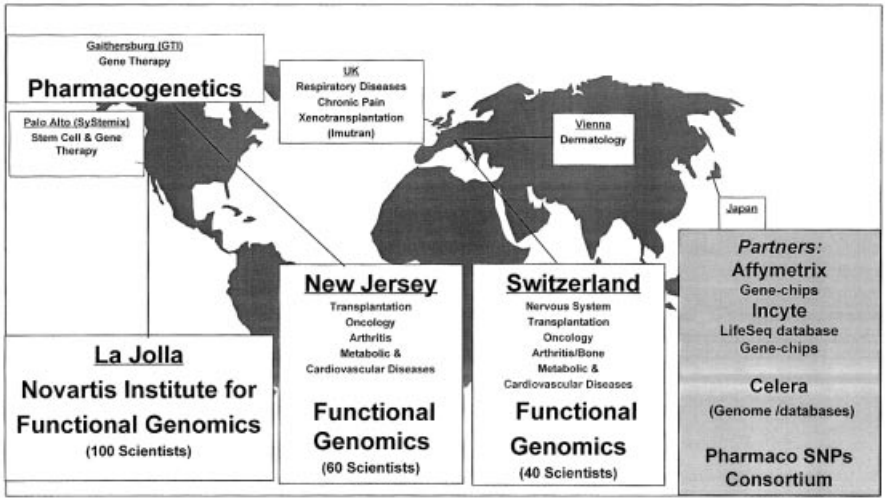


FIG. 1. Novartis pharmaceuticals functional genomics. In-house groups and major external partners.

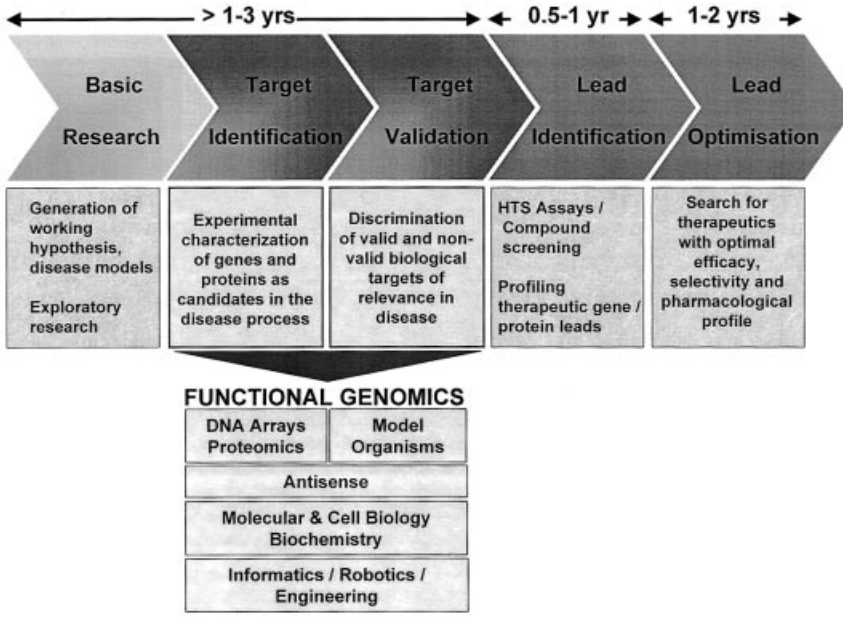


FIG. 2. Functional genomics in pharmaceutical discovery.

companies excel in one or a small subset of the necessary technologies or informatics tools. Again, this firmly underscored the requirement for Novartis to build its own in-house group in order to integrate expertise from multiple genomics approaches and technologies (Fig. 3). It also generated the core of scientific excellence necessary to judge the merit and strengths of potential external collaborators in the field offering programmes to support our in-house goals (biotech companies, industry consortia, academic groups). Novartis has ongoing external collaborations with Celera (genome information and databases), Incyte (gene-chip technology and the LifeSeq database), Affymetrix (gene-chip technology) and is a member of the Wellcome Trust/Industrial Consortium to generate a single nucleotide polymorphisms (SNPs) map.

Scientific and technology challenges in functional genomics

Given the large and increasing amount of gene sequence information available from the different genome projects, the challenge we face is the

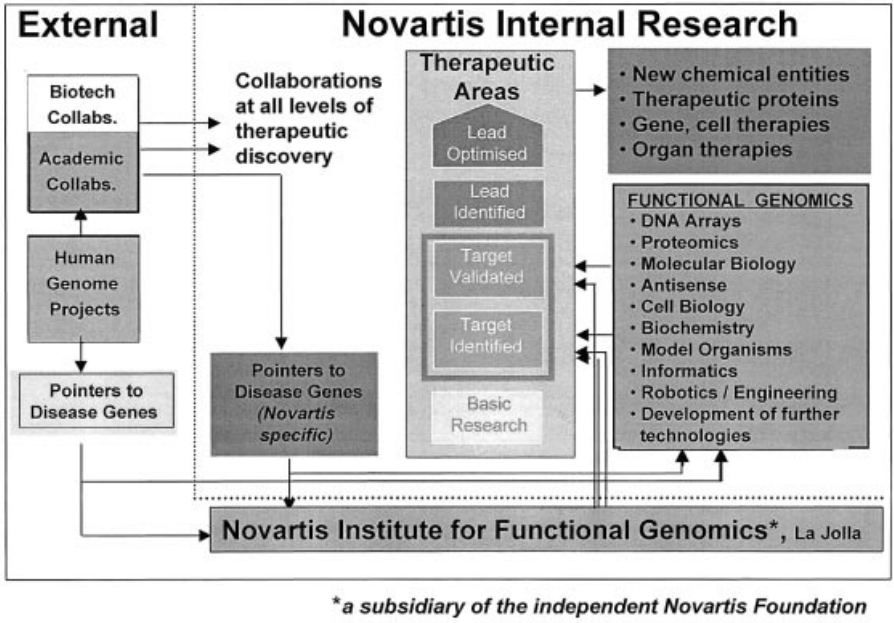


FIG. 3. Functional genomics integration in Novartis discovery research.

ability to handle the enormous amount of information resulting from analysis of differential gene expression studies. Currently, a key limiting factor in functional genomics which slows its applications is the lack of fully automated, high-throughput gene functional profiling technologies.

Functional genomics technologies

- Measurement of differential gene expression in healthy versus diseased tissues, or in drug-treated versus control cells is an important genomics approach. High-throughput measurement of relative abundance of particular mRNAs within the pool of cellular messages can be achieved using differential display approaches based on differential display, reverse transcriptase PCR and DNA array technologies (Liang & Pardee 1992, Lillie 1997, Shiue 1997).
- Proteomics (Wilkins et al 1997) enables the identification of differentially expressed proteins and study of their post-translational modifications. This approach is an important companion to gene

expression studies since there is often an insufficient correlation between the level of expression of different genes and the relative abundance of the corresponding proteins. Also, a protein and its post-translational modifications are not directly encoded for by the same gene; therefore the complete structure of an individual protein cannot be determined by reference to its gene sequence alone.

- Profiling tools to assess gene function range from transfecting cDNA expression libraries constructed in a variety of vectors (plasmids, retroviral vectors), antisense oligonucleotide libraries, ribozymes, *in situ* hybridization methodology and antisense sequences for the validation of specific genes. Model organisms such as yeast, *Caenorhabditis elegans*, *Drosophila* and the mouse represent some of the most important experimental systems available to understand gene function and are being used for *in vivo* gene profiling (Mushegian et al 1998). Furthermore, the opportunity to genetically manipulate homologous genes in these organisms will provide important contributions for identification of gene function and give valuable clues as to their potential role as a disease mediator in humans.
- Compared to development of biology-oriented technologies, less effort has been addressed to the computational biology underlying data analysis and interpretation; the establishment of so-called dry labs. This key area needs to be strengthened if the enormous amounts of genomics data are to be handled in a meaningful manner (Kingsbury 1997).

These technologies and approaches need to be used in an interactive manner in order to successfully assign gene function, place individual genes into biological pathways, predict which pathways initiate the disease process and use this information to screen and optimize therapeutic leads and candidates. It will also be necessary to accurately model dynamic interactions of cell signalling pathways to improve the prediction of effects resulting from their experimental or therapeutic manipulation. One further strength of applying genomics in the pharmaceutical arena is the extensive knowledge and the availability of *in vitro* and *in vivo* model systems to study disease pathophysiology and integration of functional genomics technologies with more established

scientific disciplines (e.g. protein chemistry, biochemistry, pharmacology, physiology).

Outlook and conclusion

Functional genomics approaches and technologies will impact on other areas of pharmaceutical R&D beyond discovery research. Pharmacogenomics activities to profile efficacy and side effects profiles of new and existing therapies in subsets of patients within a given disease category also holds the promise of addressing 'personalized' medical needs. In the molecular diagnostics field, availability of a meaningful number of precisely located SNP sites spanning the genome holds promise for association of particular genetic loci with major disease states (Venter et al 1998). This information, together with the high-throughput gene-chip technologies will offer new opportunities for molecular diagnostics and disease predisposition monitoring in large sections of the population. It will also allow much earlier preventive treatment in many slowly evolving diseases.

In conclusion, the genomics revolution is now entering its second phase whereby the pioneering efforts to map and sequence the human genome, and the enormous wealth of data they have generated, are now being converted into precise information on gene and protein function in normal and disease states. The progress of functional genomics will focus pharmaceutical research towards disease relevant targets and provide a starting point for discovery of causal and disease modifying therapies to address society's most outstanding medical needs.

References

Cohen D, Chumakov I, Weissenbach J 1993 A first-generation physical map of the human genome. Nature 366:698–701

Dyer MR, Cohen D, Herrling PL 1999a Functional genomics: from gene to new therapies. Drug Discov Today 4:109–114

Dyer MR, Cohen D, Herrling PL 1999b Building a competitive functional genomics platform. Nat Biotechnol (suppl) 17:BE18–BE19

Fields C, Adams MD, White O, Venter JC 1994 How many genes in the human genome? Nat Genet 7:345–346

Goedert M 1997 Familial Parkinson's disease. The awakening of alpha-synuclein. Nature 388:232–233

Kahn CR, Vincent D, Doria A 1996 Genetics of non-insulin dependent (Type II) diabetes mellitus. Annu Rev Med 47:509–531

Kingsbury DT 1997 Bioinformatics in drug discovery. Drug Dev Res 41:120–128
Liang P, Pardee AB 1992 Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 257:967–971
Lillie J 1997 Probing the genome for new drugs and targets with DNA arrays. Drug Dev Res 41:160–172
Maddison PJ, Isenberg DA, Woo P, Glass DN (eds) 1993 Oxford textbook of rheumatology, vol 2. Oxford University Press, Oxford
Mushegian AR, Garey JR, Martin J, Liu LX 1998 Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes. Genome Res 8:590–598
Polymeropoulos MH, Lavedan C, Leroy E et al 1997 Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. Science 276:2045–2047
Postma DS, Bleecker ER, Amelung PJ et al 1995 Genetic susceptibility to asthma—bronchial hyperresponsiveness coinherited with a major gene for atopy. N Engl J Med 333:894–900
Selkoe DJ 1997 Alzheimer's disease: genotypes, phenotypes and treatments. Science 275:630–631
Shiue L 1997 Identification of candidate genes for drug discovery by differential display. Drug Dev Res 41:142–159
Venter JC, Adams MD, Sutton GG, Kerlavage AR, Smith HO, Hunkapiller M 1998 Shotgun sequencing of the human genome. Science 280:1540–1542
Watson JD 1990 The human genome project: past, present and future. Science 248:44–49
Wilkins MR, Williams KL, Appel RD, Hochstrasser DF (eds) 1997 Proteome research: new frontiers in functional genomics. Springer Verlag, Berlin
Yankner BA 1996 Mechanisms of neuronal degeneration in Alzheimer's disease. Neuron 16:921–932

DISCUSSION

Venter: It's hard to disagree with your strategy, but like anything, whether it works will be determined by whether many of the details of implementing it work.

Herrling: The question really concerns the assumption that putting so many resources into genomics will give us a competitive advantage. Everybody is doing something with genomics, but the magnitude of our effort and having a large number of therapeutic targets is what we think may give us such an advantage. If it doesn't work, then it is a big waste of money.

Venter: What sort of proportion of your R&D budget are we talking about?

Herrling: It is 10%. But this does not count the people in the therapeutic areas themselves, which is the second part of the equation.

Venter: What was the driving force behind the change in direction at Novartis?

Herrling: Our experience with biotech. For many of our biotech projects, we have found that for one product we needed multiple licences. With the fact that the patents have moved away from the classical chemical patenting, where in the end the patent scope is reduced to what is documented by examples, instead they have become so broad that everybody is interfering with everybody else. It is a complete mess, and the only people making money will soon be the lawyers.

Straus: Not the academic ones!

Herrling: What is happening now is that everybody has pieces of every protein. If you have nothing and everybody else has everything, you are in a bad situation. We are now trying to patent defensively in order to be able to maintain freedom of operation and to trade.

Venter: Those lessons have been learned the hard way in the agribusiness area.

Herrling: We don't want to wake up that late!

Venter: In the agricultural side of this industry, patents are an absolute determination of what you can do. The agbiotech industry is absolutely ruled by who has intellectual property.

Herrling: A second aspect that is important is that as long as people ask simply for a fee for using a tool that is fine. But when they ask about reach-through payment of royalties, this is a real problem. It is as if IBM were asking for royalties on all the drugs that we analyse with their PCs. This is not reasonable, but some people are actually suing us in order to push these kinds of things through.

Straus: I raised earlier the question of collective rights versus exclusive rights. I am not a protagonist of that. However, hearing about your experience where for example one party has a patent for the promoter, another has a patent for the 3′ untranslated region, another has a patent for the vector and another for the gene, you need 10 or more licences for one single product and even if you behave cautiously it seems that you can get a licence from someone who has a poor patent and you are blocked altogether. Is there anything which should be discussed seriously with regard to this? Would it make sense to attempt a system similar to that we have in the field of copyright, where for many uses only a right to collect royalties exists, not however an exclusive right.

Venter: If everybody does defensive patenting, isn't it almost better not to have any patents at all?

Herrling: I think that we will be forced to reform the patent laws.

Straus: It could damage a valuable system, in the end.

Lipshutz: The electronics industry suffered this problem in the early days. There was very extensive cross-licensing, with 500–600 patents at a time being cross-licensed. People get themselves in exactly the position you describe: a company will go out and just start inventing just for the purposes of building up a portfolio in an area so that they can compete.

Herrling: The other solution would be to try to narrow down the patents to cover the smallest possible pieces, rather than going for the biggest possible scope. In chemistry there are no such problems: patenting molecules and compounds is a fairly straightforward process.

Straus: Courts and offices should bear in mind that the scope of protection should always be commensurate with the contribution to the art. You should get solid protection for your contribution but not for more.

Venter: I disagree. The person who digs up a new antibiotic from the dirt in their backyard should get a patent that recognizes the value of discovery of the invention, not the amount of sweat that went into it.

Straus: Don't misunderstand me: I'm not saying that the intellectual input is not important, but it is the importance of the contribution that is relevant, objectively viewed.

Goodfellow: Just a comment about everyone being happy about patenting compounds: the truth is, we are only happy with this because we have been doing it for 30 years, and we are used to it. In 10 years time everyone will be comfortable with patenting DNA sequences.

Venter: The combinatorial approaches are creating new problems in chemical patents. This is true for many technologies.

Roses: The basic philosophical problem is that there has been a biotechnology industry that has put value in creating interference. Biotech has a very different value system from the pharmaceutical industry. Because defence is often the best form of offence, pharmaceutical companies can also make it difficult by creating interference by blocking everybody else out. This doesn't mean that those things can't be cross-licensed by other pharmas, because the other part of the pharmaceutical industry, which I didn't know about until I joined, is that compounds get switched and bought all the time for other reasons. The sort of patenting we do is

defensive against an industry that has been built up to attack profit margins and raise the cost of drug development.

Venter: But why is it OK for GlaxoWellcome to do but not the biotech company that invested their equity and knowledge?

Roses: What I'm saying is that we are just doing the same thing.

Venter: But on a much broader scale.

Roses: Yes.

Mann: With gene patents, is this a problem that is time-limited, in the sense that so many people have now filed patents, for example, on secreted proteins, that they are virtually all claimed. Do we only have to wait a number of years and all research on these molecules will be free?

Straus: Patents are always time-limited, although most people would then try to find a way to get some additional patent protection. In principle a patent protection expires after 20 years.

Herrling: That's is an important point. I think many people are not aware that patenting as soon as they can is not necessarily a good idea, because biotech product development is not a rapid process. By the time your drug is on the market, you may only have a few years of patent protection left.

Roses: Patenting is publishing. If you are in the business to make money out of patents in the short-term, as are the venture capitalists, then you don't care about the long-term.

Straus: But it is a race. If you are second, you are a loser, no matter how good your product is.

Goodfellow: I think we forget that the patent system is a great system. There is a window of opportunity for recovery of investment, and then at the end of 20 years the drugs can be produced and sold anywhere around the world by anyone. The commercial drive is always there, so it is the job of lawyers to try to find ways to extend patents, but at the end of the day the patent expires.

Venter: There is a danger that with all these mergers we end up with just one big pharmaceutical company.

Roses: Large pharmaceutical companies all think in the same way. They are going to get these patent estates. They are going to trade them in just the same way that they trade compounds and everything else: they are looking at the end-product of their business, they are not necessarily looking at interference.

Herrling: Not at all. It brings no money in.

Venter: So should Celera (or whoever gets there first) be patenting the human genome? Is the goal over the next five years to build the biggest intellectual property estates for trading?

Herrling: Actually, the goal is not to build the biggest estate, but the right one. Patents cost a lot of money, and you can't simply patent everything.

Venter: The patent commissioner told us that the largest patent application they recently got was 250 000 pages long, and they have had several of over 100 000 pages from Incyte and HGS.

Herrling: That's good; it will take 10 years to read them!

Venter: But these would be small compared to provisional filing applications for single nucleotide polymorphisms (SNPs) from the SNP consortium. Each SNP has 100 base pairs per SNP with 300 000 SNPs. There is a lot of filing here, even if the goal is just to get a statutory invention registration.

Herrling: How do you see these patents evolving? There must be some sort of limit.

Straus: We had this discussion in the Standing Advisory Committee of the European Patent Office. There are different approaches. One is fees: page fees and claim fees. Perhaps a better approach is an approach as to substance, such as preventing people claiming inventions which, like SNPs or expressed sequence tags (ESTs), are trying to cover extremely broad areas. For example in the Incyte patent one can read: 'The subject invention provides unique polynucleotides (SEQ ID 1–44) which have been identified as novel human Kinases (Kin). These partial cDNA were identified among the polynucleotides which comprise various Incyte cDNA libraries. The invention comprises polynucleotides which are complementary to the kinase sequences (SEQ ID Nos 1–44). The invention comprises also the use of Kin sequences to identify full-length human kinases…' and so forth. I don't know where the contribution really is.

Venter: As you know, I have been quoted as saying many times that in the DNA patent applications, the main inventor is the patent attorney, because that's where all these claims come from. Why is that any different than patenting combinatorial libraries?

Straus: You should have a better answer than I. I have the feeling that in this area things are much more interdependent, much more interrelated than in the classical synthetic chemistry areas.

Venter: Is that true? In chemistry, it seems that there is only a finite amount of chemical space.

Goodfellow: No, there is an infinite amount of chemical space, but the chemical space which is compatible with producing drugs is more constrained.

Rubin: Is there a limit to how many sequences that can be put in one patent?

Venter: The answer depends on whether they are independent or linked somehow. You can only file 10 at once in the USA.

Rubin: If you can only file 10 human genes in each patent, there is a strong economic incentive not to patent every gene but only those you have an interest in.

Venter: The approach that several groups and companies have taken with microbial genomes is to patent the entire genome as one sequence.

Goodfellow: The real question is what is the utility? In today's world, a patent has to issue before you stop anybody else screening. So if it's something which is interesting and is obvious, then the chances are that you will get there just as quickly as the person who patented it, and you'll have a window of opportunity when you can screen. If you miss this opportunity, you may still be able to screen, because you don't necessarily have to use the recombinant product in order to do a screen. You could search for a cell line which naturally expresses the molecule that you're interested in. In many cases, the reason for patenting a target may be very defensive.

Venter: I'm not sure that is true for the agribusiness industry right now.

Goodfellow: Another approach is to patent a mode of action. If you know enough about a target to understand the biology, then it's possible that you could patent that information. That is, a drug which inhibits this target will have the following effect. This requires you to have done the biology to know what the effect is going to be, at which point we have come all the way round the circle.

And the difference really is this. There used to be a time when we only had 400 targets. Academics created most of the biology to go with those targets. Today, companies are making big investments creating the biology to go with the targets. They want to protect that investment they have made in the biology.

Venter: If the US Patent Office issues patents with extremely narrow scopes, as I understand they intend to, then nobody is ever going to be affected by the Incyte patent Joseph Straus mentioned earlier because it will cover just those sequences as they describe them. I hope that's the way it's going to go, because it will allow the best of both systems. You get a patent on what you actually did discover or invent, but it doesn't reach out to cover things beyond the sequence in the filing. The typical practice now is to claim all genetic variations in the population. For instance, when Francis Collins filed the patent on the cystic fibrosis chloride ion channel gene, if it covers the use of that gene for detecting cystic fibrosis that's fine, but if it covers all of the other variations that ever occur in the human population forever, that is more troubling.

Straus: We should rely on the courts to rule on scope, but the problem is that the costs of litigation are extremely high.

Venter: Also, the courts move at glacial speed. The field has changed already again before the issues created with the ESTs at the beginning of this decade have had a chance to go through the courts. ESTs aren't really an issue anymore and the courts still can't even work out what they're going to do with them. There is a disconnect between the pace at which the law works and the pace of change in science.

Lipshutz: Litigation is something which we absolutely try to avoid. We only turn to it as a last resort. It is not just the financial aspect, it is also a great distraction.

Venter: It's cheaper to take half the price of the licence.

Straus: Yes, but you may not mix up companies like yours, and where you have real competitors. Real competitors are not interested in granting licences. As a rule, they want the market for themselves.

Venter: So how does the compulsory cross licensing work, or doesn't it?

Herrling: We had a case during the merger with our gene therapy patents, where one of the parts of the deal was that we had to give licences to whoever wanted them.

Straus: This was because the original invention was owned by the NIH.

Herrling: We had an exclusive licence on that. The real issue is that they were afraid that this was a monopoly position.

Venter: I thought that was the point of a patent!

Straus: The US is always fighting against compulsory licenses. However, in practice the Administration and Courts seldom arrive at similar results, i.e. as if there were a compulsory license rule. Under conditions set forth in the International Agreement on Trade Related Aspects of Intellectual Property Rights (TRIPs), in most European countries compulsory licences are available in case of dependent patents.

Venter: Does it work?

Straus: It works in part just by its existence, because you know in the end you could be forced, so you may be more willing to negotiate.

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

Final discussion

Venter: Large-scale genome sequencing has changed the basis of our science in ways that I don't think any of us anticipated. Excellent examples of this were given at this symposium with the papers on proteomics, where a little bit of peptide sequence going against expressed sequence tag and genomic libraries has allowed the protein world to expand exponentially—and sooner than was anticipated. Another example is the AFLP technology: with just a one or two base extension on a restriction enzyme site, going back to the databases you can get expression and other data very quickly.

Goodfellow: Bacterial genome sequences have completely changed the probability of whether we will have new antibiotics before we have an even more serious antibiotic resistance problem. Having complete genome sequences makes it possible to take a genomic approach towards the problem of identifying targets.

Venter: In this regard, it is worth noting that out of the 500 or so pharmaceutical targets we had before genomics, less than 20 were being targeted for antibiotic development.

Goodfellow: In my mind, what has happened with the bacterial genomes is a paradigm for what we hope will happen when we have the complete human genome sequence.

Venter: It is not going to be too long time before the results of microbial genomics begins to show up in the clinic. Tremendous work has been done by Marty Rosenberg's group at SmithKline Beecham, for example.

Many of the new vaccines weren't driven by genomics, but hopefully the DNA vaccines driven by having the sequence will change the vaccine paradigm also. However, the AIDS genome has been around for a long time, so if that's a paradigm for how genomics will drive things, it's impact will be limited. But it is still a forward step having all the genes.

Goodfellow: You say that, but actually we could put it round the other way. Where would we be if it wasn't possible to rapidly sequence the

genomes of viruses? Clearly, you wouldn't think of working on a virus today until you had the sequence of the virus in front of you.

Venter: Do any one of our three proteomic experts want to predict where proteomics will be in five years?

Hochstrasser: I believe that the number of proteins you can identify thanks to the genome will increase dramatically, through high-throughput technology. I believe also that scanning technology will be extremely powerful. One current gap, however, is 3D crystallography, which is pretty slow. And going from the final 3D structure to the function is still a problem.

Venter: That's a good point. There are attempts to scale this up around the world.

Goodfellow: I wrote a 'News and views' for *Nature* in 1992, in which I made a series of predictions of when we would complete the sequence of the human genome (Goodfellow 1992). If you start when Sanger invented DNA sequencing, and look at the accumulation in the sequence which is available in the public databases, for the first 10 years it doubled every 18 months. And about 10 years ago it went to doubling every year. Continuing that, I made the prediction that the human genome would be sequenced on 1 January in the year 2000, which is not going to be that far out.

If you do exactly the same extrapolation for 3D structures, and ask when we will have the 3D structure of the 100 000 proteins encoded by the human genome, the answer comes out in the year 2014. Even if things just happen like they're happening now, it's not going to be too long before we have the whole human genome's worth of 3D structures.

van Oostrum: That would imply that the massive amount of money now going into the human genome sequencing should also be allocated to structural genomics.

Goodfellow: It doesn't imply anything of the sort. When we started off sequencing the human genome, it was argued that it would cost US$3 billion. I remember getting up and arguing in front of politicians that instead of buying one Trident nuclear submarine, perhaps the British government should invest the same amount of money in sequencing the human genome. Actually, the cost hasn't gone up like that—the cost has gone up in a linear fashion, while the sequence accumulation is going up exponentially.

Venter: I was at the meeting where that figure of US$3 billion was derived. It was a highly arbitrary process: someone asked about how many bases there are in the genome. People guessed around three billion (we still don't know within 25% either way on the total). Someone then asked Lee Hood what he thought sequencing costs could possibly come down to, and he said about a dollar a base pair, so we asked for three billion dollars. If you calculated the cost of getting 100 000 crystal structures, US$3 billion would look trivial, but the costs are changing very rapidly. We have tried to set up a mini high-throughput facility in Maryland, where The Institute for Genomic Research (TIGR) has been purifying and crystallizing proteins. It is amazing that genomes like *Methanococcus*, a thermophilic organism, are actually providing rapid crystal structures, because for some reason the same properties that allow the proteins to survive high temperatures allow them to form nice crystals fast. There are all kinds of little changes that could result in the same rapid scale-up of getting 3D crystals. The hardest thing is predicting the future, particularly with the rate of change of technology.

Goodfellow: There's actually been no change in DNA sequencing techniques.

Venter: It is true that we are still using Sanger sequencing, but the instruments for doing Sanger sequencing technology have changed substantially.

Goodfellow: There has been an automation of the process we already had, and I would argue the same will be true for crystallography. We have already seen the partial automation of this process, but the technology is there, it is robust and it works. If you could come up with a more efficient way of crystallizing proteins, then the problem is largely solved.

Hochstrasser: If the companies deliver, in a year's time we will be able to identify 1000 proteins a day from a single 2D PAGE.

Venter: Will we have automated protein readers? I read an article on the plane coming over here about this automated house in Japan, where they have this fancy toilet that determines your weight, your fat body mass, measures the glucose automatically in your urine, and looks for occult blood! There is a radio transmitter that will then immediately send data from your toilet to your physician. Are we going to have a GC mass

spectrometer hooked up to everyone's toilets, right into the databases? Gerry Rubin, are you going to volunteer for that?

Rubin: I was going to volunteer to change the subject.

From my point of view the mass spectrometer is going to change protein biochemistry, because defining this 'periodic table' of the universe of proteins will then make the problem of protein identification much easier. Many of the experiments that have already been done, where people have isolated complexes of say 50 proteins but can only identify three of them, will be re-done. We will then get a much better idea of the organization of proteins within the cell. For people who work on small animals, the high sensitivity of this technique is going to be very important, because it takes away much of the burden of only having small amounts of material.

Mann: I suspect that people will use these techniques to get a truer picture of the actual structure of the cell and also of changes in proteins that happen not through transcription but by other regulatory mechanisms on a very fast time-scale and which we had previously no way of measuring.

Venter: Another area we have concentrated tremendously on is model organisms—we have covered microbes, yeast, *Caenorhabditis elegans*, *Drosophila*, mouse, *Arabidopsis* and humans. I think we have seen that there is no single model organism that is sufficient, including human. For example, I'm not sure that many people would have predicted a few years ago that studying DNA repair in *Escherichia coli* would help us to understand colon cancer. The message is that there's no wrong species to look at.

Goodfellow: In the real world, anybody in this room who gets a sequence, goes and looks in every other database to see whether that sequence has got any annotations associated with it. If you find a related sequence in yeast you're pleased because you know it's going to give you a very easy route to asking something about the biochemistry.

Venter: We have the databases in the worm, fly and yeast biology. It seems like those indexes will be absolutely invaluable in terms of the starting point.

Rubin: A philosophical point about genetic screens. These were the first real whole genome experiments. For example, when you do a saturation screen in *Drosophila* you are looking in an unbiased way for all the genes in

the organism that can generate a given phenotype. There have been large numbers of successes, like looking at the cell cycle in yeast or embryonic patterning in *Drosophila* where people have actually done that. This is one of the few really proven approaches in functional genomics. The availability of genomic sequences and maps has dramatically improved our ability to do positional cloning. It used to be that you did your screen and then it would take 20 person years to clone the genes. Shortening that time from 20 years to one year is going to make the forward genetic approach much more powerful. If we have enough people doing genetic screens, they will generate a tremendous amount of annotation. I believe that the way we'll figure out the function of the majority of human genes is by doing model organism genetics.

Hochstrasser: Think about the gap between the medicine we have today and the medicine we will have tomorrow. It is like the difference between medicine before and after radiology. An immediate impact is found in the microbiology lab.

Goodfellow: If you have the complete sequence of every human being, can you use that sequence information to deduce things about the control and timing of gene expression, and so on? As a hard-line reductionist, I believe that all the relevant information is in the DNA and we don't know how to read the code. But if we have lots of codes available then maybe we will be able to extract that information.

I think that you have to define the data sets that you want to collect, on all genes. We take the lowest-hanging fruit first, so the easy question is to ask where the genes are expressed. Next might be the construction of a library of knockout mutants. The experiments which were done at Amgen are just helping to define the sorts of answers that you would like. I would generalize these and then collect the data on all genes in the most effective, cheapest way.

Venter: In genomics, I strongly believe the most powerful tool will be the computer and informatics, because we don't have sufficient knowledge base to use any other assessment, particularly when it comes to regulators and modifiers and the timing. This is why we feel the human genome is almost worthless for most practical purposes. Even with mouse it is going to be difficult to add the interpretation. It would be nice to have every mammal for that reason. When people at TIGR and other places line up a couple of dozen microbial genomes, the

comparative information will be incredibly powerful in terms of sorting out the broader pieces of evolution and the biology.

Goodfellow: There's a circle, which we haven't talked about. That is, the information that you generate by studying expression should itself be interpretable in terms of the DNA sequence. If you know all the genes in yeast which are induced in the presence of glucose, or you know all the genes which are affected by growing anaerobically, that ought to provide clues to sequences which control gene expression. Eventually, when the circle is complete we can start making predictions: 'This gene has sequences in front of it which suggests that it should be expressed under the following conditions; let's go and look on the databases to see if that hypothesis is true.'

Venter: I liken this era to the situation roughly 100 years ago with descriptive anatomy. We are in a descriptive phase of biology, which the major funding institutions have decided that we passed a long time ago, and now you can only get a grant in the USA if it's hypothesis-driven, except for a limited number of these sequencing projects. How do we move forward descriptive biology, until we get to the point where that can generate some hypotheses? These programs are being improvised at Stanford. TIGR has used its own endowment to do this and a lot of work has been funded by private industry. In the scientific community in the USA or Europe, there is not a means for doing descriptive biology.

Goodfellow: This is because we all grew up in the same system, and we all want to grow up to be heads of the lab like the head of the lab that we trained in.

Rubin: I like Pat Brown's term 'hypothesis-limited research'. Largely because of the increase in the NIH budget, the realization has developed that giving more hypothesis-driven RO1 grants is not necessarily the best way to spend the money. Large data collecting efforts are beginning, but it took a long time and probably would never have happened if the budget hadn't risen and policy-makers had to look for innovative ways of spending the money.

Venter: I can't tell you how many pathogen sequencing grants TIGR had returned with comments saying that it was not clear what the hypothesis was, including a project to sequence *Staphylococcus aureus*, which had the obvious hypothesis that if we had the sequence it might drive forward new therapeutics.

Rubin: It goes further back. For a long time the genetic studies section of the NIH would not give a grant for any genetic screen, because these efforts were considered fishing expeditions and not hypothesis-driven.

Venter: Do you think it has changed?

Rubin: It is beginning to change.

Fraser: I agree with Gerry Rubin. My experience is that now that this information has been percolating through the system for a few years, funders are coming to the realization that to fully exploit the power of all this information we have to get away from the traditional ways of funding science. There is a tremendous amount to be gained by just going in and doing what in the past would have been considered fishing expeditions.

Venter: Between Peter Goodfellow, Paul Herrling, Allen Roses and Larry Souza, we have a group of leaders in the pharmaceutical industry that clearly believe in the applications of genomics. These are representing four companies that are clearly putting their money where their mouths are. But it is not a universal approach yet in the pharmaceutical industry.

Goodfellow: There is certainly a difference in emphasis, but very few companies are not signed up to one of the databases from Celera, Incyte or HGS. There are only one or two exceptions.

Venter: There are one or two noted exceptions. One is what was not too long ago, the largest pharmaceutical company in the world, before mergermania. They clearly have an anti-genomic approach. Can a company survive without incorporating genomics in the future?

Goodfellow: Of course. There are lots of different ways of making a company. You can make a model where you don't do any target discovery. I can give you a list of good drug targets, which all of us would agree about, to which there is no drug currently available. If you have new chemistry technology, they would be perfectly acceptable targets. I could also give you a list of top-selling drugs, even to this day, for which the targets are unknown. You could make a very nice living by finding out what those targets are and then making improved drugs against them. There are many strategies for making a company. But I have to agree: you can't be a major pharmaceutical company without genomics.

Venter: The last topic I was going include in my summation is the discussion we had on the individual genetic variation. I think that we

will be doing gene and target discovery for the next century, in terms of finding out that that a particular gene is actually linked somehow to a particular trait or disease. The question is, how can we improve the limited population of therapeutic efficacy? Is individualized genetic variation going to be the key to that? This gives us a role for the patent attorneys and ethicists in terms of the societal implications. I'm very concerned that we don't go back to what happened to Cold Spring Harbor in the 1930s and develop the 'new eugenics', from a combination of sloppy science and the kind of news sensationalism that led to the 'headless human' story that I mentioned earlier. The same clinical genotypes that Allen Roses and others would like to do, to select clinical populations, are the subject of a great social debate. There have been huge debates on the genetics of violence, people want to find genetic links to all human traits, habits and skills. The scenario that I fear is an extension of the laws in the USA requiring the police to tell people in the neighbourhood when a former child molester moves into their neighbourhood after release from jail. It doesn't seem so far-fetched to me that people might want to genotype all child molesters in prison, and screen the population for any genetic links to this. All these scenarios are based on a belief in genetic determinism which was the standard in biology not so long ago, and unfortunately it is still part of the public and press interpretation of science.

Herrling: One aspect related to that. If a proportion of information will be found at the DNA level about responders and non-responders, what will the proportion be that will not be seen at the DNA level? If there subtle changes or individual differences that exist at the protein interaction level, what is seen at the DNA level may be very limited.

Venter: Some things will be very clear-cut: if you have an amino acid change in the receptor binding site that determines whether you respond to the drug or not, that's going to be largely predictable.

Goodfellow: This is the problem, and it is why I was being a little negative, in a 'tongue-in-cheek' way, about pharmacogenetics. History teaches us that when you look at these phenotypes, some of them will turn out to be controlled by only one or two genes. There will be cases where you will find single gene effects which have a marked effect on drug responses. We know that you can get very marked pharmacogenetic effects due to changes in metabolism of drugs which are on the market

today. Actually, if they were put on the market today rather than 10 years ago, I suspect that you would have to test for the genetic variant before prescribing the drug. But in most cases there's going to be more than one gene involved. In human genetics it has never been possible to identify two genes linked to a disease in one experiment. Then you say, 'What if there are three or four genes?'. You just have to do the power calculations. If the phenotypes we are interested in involve more than two genes, then you can forget it. I do not think we are going to find them by genetic approaches in outbred populations. It is different in inbred mice.

Venter: As an example, for hypertension it's estimated that there are 300 genes now known to be involved in controlling blood pressure.

Goodfellow: How often do you think that we will be able to identify the genetic 'cause' of response versus non-response?

Herrling: Rarely.

Venter: But that is not what these screens are looking for. They are looking for a pattern, which will be sufficient.

Goodfellow: It will only be sufficient if it predicts whether you respond or not.

Venter: It's a predictive pattern, I agree, but the scientific basis of understanding the alleles that are associated with response or non-response is not needed.

Herrling: Let's do the experiment!

Venter: I would say that this is an excellent last word. In drawing this meeting to a close, I would like on behalf of all of us to thank our hosts, the Novartis Foundation, for bringing us all together, and all the participants for contributing to the enjoyable discussion over the last few days.

Reference

Goodfellow PN 1992 Variation is now the theme. Nature 359:777–778

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

Index of contributors

Non-participating co-authors are indicated by asterisks. Entries in bold indicate papers; other entries refer to discussion contributions.

Novartis 229: From Genome to Therapy: Integrating New Technologies with Drug Development.

Print ISBN 0-471-62744-5 eISBN 0-470-84664-X

Subject index