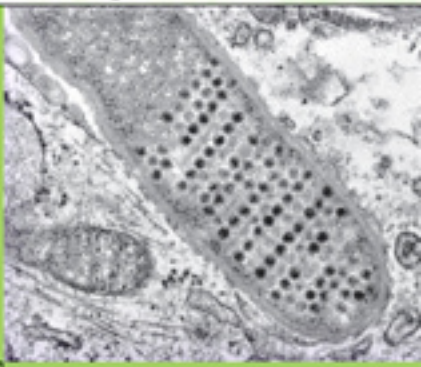
ADVANCES IN MOLECULAR AND CELLULAR MICROBIOLOGY

AMCM

Horizontal Gene Transfer in the Evolution of Pathogenesis



Edited by
Michael Hensel
Herbert Schmidt

CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521862974

This page intentionally left blank

Horizontal Gene Transfer in the Evolution of Pathogenesis

Horizontal gene transfer is a major driving force in the evolution of many bacterial pathogens. The development of high-throughput sequencing tools and more sophisticated genomic and proteomic techniques in recent years has resulted in a better understanding of this phenomenon. Written by leading experts in the field, this edited volume is aimed at graduate students and researchers and provides an overview of current knowledge relating to the evolution of microbial pathogenicity. This volume provides an overview of the mechanisms and biological consequences of the genome rearrangements resulting from horizontal gene transfer, in both prokaryotes and eukaryotes, as well as overviews of the key mobile genetic elements involved. Subsequent chapters focus on paradigms for the evolution of important bacterial pathogens, including *Salmonella enterica, Streptococcus pneumoniae*, and *Staphylococcus aureus*. The influence of socioeconomic parameters in the dissemination of transferable elements, such as antibiotic-resistant genes in bacteria, is also discussed.

Michael Hensel is currently Professor of Microbiology and Immunology at the University of Erlangen in Germany.

Herbert Schmidt is currently Professor of Food Microbiology at the University of Hohenheim in Germany.

ADVANCES IN MOLECULAR AND CELLULAR MICROBIOLOGY

AMCM

Over the past decade, the rapid development of an array of techniques in the fields of cellular and molecular biology has transformed whole areas of research across the biological sciences. Microbiology has perhaps been influenced most of all. Our understanding of microbial diversity and evolutionary biology and of how pathogenic bacteria and viruses interact with their animal and plant hosts at the molecular level, for example, has been revolutionized. Perhaps the most exciting recent advance in microbiology is the fusion of classical microbiology, microbial molecular biology, and eukaryotic cellular microbiology. Cellular microbiology is revealing how pathogenic bacteria interact with host cells in what is turning out to be a complex evolutionary battle of competing gene products. Molecular and cellular biology are no longer discrete subject areas but vital tools and an integrated part of current microbiological research. As part of this revolution in molecular biology, the genomes of a growing number of pathogenic and model bacteria have been fully sequenced, with immense implications for our future understanding of microorganisms at the molecular level.

Advances in Molecular and Cellular Microbiology is a series edited by researchers active in these exciting and rapidly expanding fields. Each volume will focus on a particular aspect of cellular or molecular microbiology and will provide an overview of the area, as well as examine current research. This series will enable graduate students and researchers to keep up with the rapidly diversifying literature in current microbiological research.

Published Titles

1. *Bacterial Adhesion to Host Tissues: Mechanisms and Consequences*. Edited by Michael Wilson 9780521801072
2. *Bacterial Evasion of Host Immune Responses*. Edited by Brian Henderson and Petra Oyston 9780521801737
3. *Dormancy and Low Growth States in Microbial Diseases*. Edited by Anthony Coates 9780521809405
4. *Susceptibility to Infectious Diseases: The Importance of Host Genetics*. Edited by Richard Bellamy 9780521815253
5. *Bacterial Invasion of Host Cells*. Edited by Richard Lamont 9780521809542
6. *Mammalian Host Defense Peptides*. Edited by Deirdre A. Devine and Robert E. W. Hancock 9780521822206
7. *Bacterial Protein Toxins: Role in the Interference with Cell Growth Regulation*. Edited by Alistair Lax 9780521820912X
8. *The Dynamic Bacterial Genome*. Edited by Peter Mullany 9780521821575
9. *Salmonella Infections: Clinical, Immunological and Molecular Aspects*. Edited by Pietro Mastroeni and Duncan Maskell 9780521835046
10. *The Influence of Cooperative Bacteria on Animal Host Biology*. Edited by Margaret J. McFall-Ngai, Brian Henderson, and Edward G. Ruby 9780521834650
11. *Bacterial Cell-to-Cell Communication: Role in Virulence and Pathogenesis*. Edited by Donald R. Demuth and Richard Lamont 9780521846387
12. *Phagocytosis of Bacteria and Bacterial Pathogenicity*. Edited by Joel D. Ernst and Olle Stendahl 9780521845694
13. *Bacterial-Epithelial Cell Cross-Talk: Molecular Mechanisms in Pathogenesis*. Edited by Beth A. McCormick 9780521852449
14. *Dendritic Cell Interactions with Bacteria*. Edited by Maria Rescigno 9780521855860
15. *Bacteriophage Ecology: Population Growth, Evolution, and Impact of Bacterial Viruses*. Edited by Stephen T. Abedon 9780521858458

Advances in Molecular and Cellular Microbiology 16

Horizontal Gene Transfer in the Evolution of Pathogenesis

EDITED BY
MICHAEL HENSEL
University of Erlangen, Germany

HERBERT SCHMIDT
University of Hohenheim, Germany

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK
Published in the United States of America by Cambridge University Press, New York
www.cambridge.org
Information on this title: www.cambridge.org/9780521862974

First published in print format 2008

ISBN-13 978-0-511-40870-0 eBook (EBL)

ISBN-13 978-0-521-86297-4 hardback

We dedicate this book to our mentors
Jürgen Heesemann
and
Helge Karch

Contents

Preface

During the past 25 years, a nearly exponential increase has occurred in nucleotide sequences available from databases. The first microbial genomes were published in 1995 (Fleischmann *et al.*, 1995; Fraser *et al.*, 1995; Himmelreich *et al.*, 1996); the NCBI database currently contains 534 complete bacterial chromosome sequences. The genomes have been determined not only from different species, but also from different strains of the same species. This has paved the way for comparative genomics and has allowed detailed analysis of genetic differences between strains (Wren, 2000; Dobrindt and Hacker, 2001; Edwards, Olsen, and Maloy, 2002; Raskin, Seshadri, Pukatzki, and Mekalanos, 2006). Since infectious diseases are a major health threat worldwide and range among the most frequent causes of death worldwide (WHO Health Statistics 2006), it is not surprising that the first two organisms to be sequenced were pathogenic bacteria.

During the past decade many attempts have been made to understand the molecular basis of microbial pathogenicity, and our knowledge has advanced quickly, in particular through the use of methods of cellular biology, genomics, and proteomics. Various events in the pathogenesis of microbial infections, such as adherence to and entry into human and animal host, invasion of host cells, toxin production, establishment and dissemination of bacterial populations in the host, and the role of the host immune system, have been studied in detail for many host-pathogen interactions. In most cases, the specific features that define microbial pathogenicity are encoded by mobile genetic elements such as bacteriophages, plasmids, and pathogenicity islands, and fast transferability is often facilitated by transposable elements (for reviews, see Finlay and Falkow, 1997; Low and Porter, 1978).

In particular, the sequencing of whole genomes and the development of more sophisticated bioinformatics tools have shown that the genomes of microorganisms in even a single species may vary enormously (for example, in *Staphylococcus aureus*; see Chapter 10). Improved methods for generation and analysis of sequence data, as well as application of techniques for molecular typing, such as multilocus sequence typing (MLST) and single-nucleotide polymorphism (SNP) analyses, have been used throughout the kingdoms of life and have shown that the sequence diversity between organisms, even between members of the same species, is greater than previously suspected. Microorganisms can be found in every ecological niche on earth, and this ecological speciation could occur only through the acquisition of new genetic traits.

Although horizontal gene transfer is much more elaborate in eukaryotic cells and can lead to changes in the progeny by meiosis and mitosis, recombination events also take place in bacteria following DNA uptake in many ways (Andersson, 2005; Ochman, Lawrence, and Groisman, 2000). Adaptation of bacteria to extreme environments is a consequence of changes in the genetic content and is stabilized by selective pressure. Genome changes may develop by gene loss, gene duplication, gene mutation, or acquisition of new genetic material by lateral gene transfer (LGT; Lawrence and Hendrickson, 2003; Ochman and Moran, 2001; Campbell, 2000). It has been hypothesized that between 1.6% and 32% of the genes of a given genome have been acquired by LGT (Ochman, Lawrence, and Groisman, 2000; Koonin, Makarova, and Aravind, 2001; Marri, Hao, and Golding, 2007).

The main mechanisms of DNA uptake in bacteria are conjugation, transduction, and transformation. These mechanisms must be followed by recombination events that allow the genetic traits to be inserted more or less stably into the chromosome. As Low and Porter summarized in 1978: "The term recombination can be used in a sense that it means a reassortment of series of nucleotides along nucleic acid molecules."

In this book, articles on the basics of lateral gene transfer are provided by authors who are leading scientists in the field.

In the first part, the chapters written by Jeffrey G. Lawrence and Heather Hendrickson (Chapter 1) and by Xavier Didelot and Daniel Falush (Chapter 2) provide an overview of the impact and the mechanisms of LGT as well as detailed insight into the mathematical approaches in evolutionary biology.

The second part of the book contains contributions to the role of distinct mobile genetic elements in bacterial evolution. This section contains comprehensive chapters on the role of bacteriophages as "accelerators" of bacterial evolution by Harald Brüssow (Chapter 3) and on the global

impact of bacteriophages (Chapter 4) from Roger W. Hendrix and Sherwood R. Casjens. Whereas the chapter by Hendrix and Casjens explains mechanisms of gene acquisition and moron acquisition of Gram-negative bacteria and their phages, Harald Brüssow points out the predator-prey relationship and gives a more evolutionary view of bacteriophages. A further important group of mobile genetic elements are genomic islands (GEI), in particular the subgroup of pathogenicity islands (PAI). In Chapter 5, Tobias Ölschläger and Jörg Hacker describe the structure and function of these elements and provide various examples for the impact of GEI and PAI in bacterial evolution.

The third part of the book includes a selection of paradigms for the evolution of important bacterial pathogens. Dawn L. Arnold and Robert W. Jackson provide a state-of-the art chapter on genomic islands in plant-pathogenic bacteria (Chapter 6); Sébastien Lemire, Nara Figueroa-Bossi, and Lionello Bossi focus on the contribution of prophages to virulence and diversity of *Salmonella enterica* serovars (Chapter 7). Elisabeth Carniel describes in Chapter 8 how pathogenicity and transmission in the genus *Yersinia* progressively evolved together with the gradual acquisition of foreign mobile genetic elements. Genomic and pathogenicity islands of *Streptococcus pneumoniae* are addressed in Chapter 9 by Barbara Albiger, Christel Blomberg, Jessica Dagerhamn, Staffan Normark, and Birgitta Henriques-Normark.

In Chapter 10, Richard P. Novick provides detailed information on the current knowledge on mobile genetic elements of *Staphylococcus aureus*. In the last part of this section, Wolfgang Witte sheds light on the influence of socioeconomic parameters on the dissemination of transferable elements (Chapter 11). The specific emphasis of this chapter is on the spread of antibiotic-resistant genes in bacteria.

In the last section, Jan O. Andersson gives insight into gene transfer events in eukaryotes (Chapter 12). Finally, in Chapter 13 Andrés Moya and Amparo Latorre describe the events that are involved in genome reduction of bacterial endosymbionts in the paradigmatic system of aphids and their endosymbionts *Buchneria* spp.

The information given in this book shows clearly that LTG is a driving force in the evolution of various groups of bacterial pathogens; however, LTG is not restricted to this kingdom. The tools of high-throughput sequencing and the development of more sophisticated bioinformatics and genome and proteome techniques facilitate the analysis of LGT and recombination and help us to understand the processes of diversification, reduction, and adaptation to different environments.

REFERENCES

Andersson, J. O. (2005). Lateral gene transfer in eukaryotes. *Cell Mol Life Sci*, **62**, 1182–97.

Campbell, A. M. (2000). Lateral gene transfer in prokaryotes. *Theor Popul Biol*, **57**, 71–7.

Dobrindt, U., and Hacker, J. (2001). Whole genome plasticity in pathogenic bacteria. *Curr Opin Microbiol*, **4**, 550–7.

Edwards, R. A., Olsen, G. J., and Maloy, S. R. (2002). Comparative genomics of closely related salmonellae. *Trends Microbiol*, **10**, 94–9.

Finlay, B. B., and Falkow, S. (1997). Common themes in microbial pathogenicity revisited. *Microbiol Mol Biol Rev*, **61**, 136–69.

Fleischmann, R. D., Adams, M. D., White, O., *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.

Fraser, C. M., Gocayne, J. D., White, O., *et al.* (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397–403.

Himmelreich, R., Hilbert, H., Plagens, H., *et al.* (1996). Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res*, **24**, 4420–49.

Koonin, E. V., Makarova, K. S., and Aravind, L. (2001). Horizontal gene transfer in prokaryotes: Quantification and classification. *Annu Rev Microbiol*, **55**, 709–42.

Lawrence, J. G., and Hendrickson, H. (2003). Lateral gene transfer: When will adolescence end? *Mol Microbiol*, **50**, 739–49.

Low, K. B., and Porter, D. D. (1978). Modes of gene transfer and recombination in bacteria. *Annu Rev Genet*, **12**, 249–87.

Marri, P. R., Hao, W., and Golding, G. B. (2007). The role of laterally transferred genes in adaptive evolution. *BMC Evol Biol*, **7** (Suppl 1), S8.

Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.

Ochman, H., and Moran, N. A. (2001). Genes lost and genes found: Evolution of bacterial pathogenesis and symbiosis. *Science*, **292**, 1096–9.

Raskin, D. M., Seshadri, R., Pukatzki, S. U., and Mekalanos, J. J. (2006). Bacterial genomics and pathogen evolution. *Cell*, **124**, 703–14.

Wren, B. W. (2000). Microbial genome analysis: Insights into virulence, host adaptation and evolution. *Nat Rev Genet*, **1**, 30–9.

Contributors

Barbara Albiger
Lund University
Department of Medical Microbiology
Malmö, Sweden

Jan O. Andersson
Institute of Cell and Molecular Biology,
Uppsala University, Biomedical Center,
Box 596, S-751 24
Uppsala, Sweden

Dawn L. Arnold
Centre for Research in Plant Science
University of the West of England
Bristol
UK

Christel Blomberg
Swedish Institute for Infectious Disease Control and Department of Microbiology
Tumor and Cell Biology
Karolinska Institutet
Solna, Sweden

Lionello Bossi
Centre de Génétique Moléculaire
CNRS
91198 Gif-sur-Yvette
France

Harald Brüssow
Nestlé Research Centre
26 Vers-chez-les-Blanc
CH-1000 Lausanne
Switzerland

Elisabeth Carniel
Yersinia Research Unit
Institut Pasteur
28 Rue du Dr. Roux
75724 Paris Cedex 15
France

Sherwood R. Casjens
Pathology Department
5200 Emma Eccles Jones Medical Research Building
University of Utah School of Medicine
Salt Lake City, UT 84112
USA

Jessica Dagerhamn
Swedish Institute for Infectious Disease Control and Department of Microbiology
Tumor and Cell Biology
Karolinska Institutet
Solna
Sweden

Xavier Didelot
Peter Medawar Building for Pathogen Research
South Parks Road
Oxford OX1 3SY
UK

Daniel Falush
Peter Medawar Building for Pathogen Research
South Parks Road
Oxford OX1 3SY
UK

Nara Figueroa-Bossi
Centre de Génétique Moléculaire
CNRS
91198 Gif-sur-Yvette
France

Jörg Hacker
Institut für Molekulare Infektionsbiologie
Universität Würzburg
Röntgenring 11
97070 Würzburg
Germany

Heather Hendrickson
Department of Biological Sciences
University of Pittsburgh
Pittsburgh, PA 15260
USA

Roger W. Hendrix
Pittsburgh Bacteriophage Institute and Department of Biological Sciences
A340 Langley Hall
University of Pittsburgh
Pittsburgh, PA 15260
USA

Birgitta Henriques-Normark
Swedish Institute for Infectious Disease Control and Department of Microbiology
Tumor and Cell Biology
Karolinska Institutet
Solna
Sweden

Michael Hensel
Mikrobiologisches Institut
Universitätsklinikum Erlangen
Wasserturmstr. 3-5
91054 Erlangen
Germany

Robert W. Jackson
School of Biological Sciences
University of Reading
Reading
UK

Amparo Latorre
Institut Cavanilles de Biodiversitat i Biología Evolutiva
Universitat de València
Spain

Jeffrey G. Lawrence
Department of Biological Sciences
University of Pittsburgh
Pittsburgh, PA 15260
USA

Sébastien Lemire
Centre de Génétique Moléculaire
CNRS
91198 Gif-sur-Yvette
France

Andrés Moya
Institut Cavanilles de Biodiversitat i Biología Evolutiva
Universitat de València
Spain

Staffan Normark
Swedish Institute for Infectious Disease Control and Department of Microbiology
Tumor and Cell Biology
Karolinska Institutet
Solna
Sweden

Richard P. Novick
Skirball Institute and Department of Microbiology
New York University School of Medicine
540 First Avenue
New York, NY 10016
USA

Tobias Ölschläger
Institut für Molekulare Infektionsbiologie
Universität Würzburg
Röntgenring 11
97070 Würzburg
Germany

Herbert Schmidt
Institut für Lebensmittelwissenschaft und Biotechnologie
Fachgebiet Lebensmittelmikrobiologie (150a)
Universität Hohenheim
Garbenstrasse 28
70599 Stuttgart
Germany

Wolfgang Witte
Robert Koch Institute
Wernigerode Branch
Burgstraße 37
38855 Wernigerode
Germany

Part I Theoretical Considerations on the Evolution of Bacterial Pathogens

CHAPTER 1

Genomes in Motion: Gene Transfer as a Catalyst for Genome Change

Jeffrey G. Lawrence and Heather Hendrickson

> The change from species to species is not a change involving more and more additional atomistic changes, but a complete change of the primary pattern or reaction system into a new one, which afterwards may again produce intraspecific variation by micromutation.
>
> – Richard Goldschmidt, 1940

1.1. INTRODUCTION

Despite our interest and motivation, bacteria are not particularly easy organisms to study; their niches are complex and poorly understood and the vast majority of these species are difficult to culture or to manipulate in the laboratory. Of all bacteria, it is pathogens whose physical, social, and economic impact on our day-to-day lives garners the most attention, from both scientists and non-scientists alike. As a result, pathogens are among the best-studied bacteria, and lessons we learn from them are often generalized to other, non-pathogenic bacteria. Not surprisingly, the first lessons learned in the so-called genomic era came from pathogens, which were the first organisms with fully sequenced genomes. The promise of genomics was that the limitations of conventional microbiology could be overcome by studies of genome sequences and careful analysis of the genes contained therein. Here we examine how genomics has shaped our understanding of microbial genome evolution and ask how extensible these lessons may be. Among the notions that attracted widespread attention was the finding that certain clusters of genes are specifically responsible for virulence and that these loci are often obviously of foreign origin, having been introduced by the then under-appreciated process of lateral gene transfer or LGT.

Coming more than a decade after these findings, this volume is focused on the hugely influential role of LGT in the evolution of genomes, particularly those of pathogenic bacteria. The debate over the relative importance of lateral gene transfer is as old as discussion of the phenomenon itself (Doolittle, 1999a, 1999b; Kurland, 2000). Yet most will agree that gene transfer has not only been instrumental in the origin and diversification of pathogenic bacteria and fungi (Groisman and Ochman, 1994, 1997), but that the very nature of pathogenic organisms has been shaped by this process. In a world lacking LGT, pathogens would be very different creatures from the ones we see today. But to many biologists, LGT is difficult to integrate into a conceptual framework of organismal evolution. Darwinian evolution is traditionally viewed as gradual change, with endless cycles of minute refinements shaping and adapting an organism to its environment. Here, individuals combine and reassort subtle variants within populations to make fluid and natural transitions between phenotypic states. Those changes which increase fitness within a particular environment would be retained. In contrast, change induced by lateral transfer can be both startlingly quick and strikingly large; LGT can catalyze the sort of dramatic phenotypic change that recalls discussion of "hopeful monsters" in the middle part of the last century (Goldschmidt, 1940) wherein speciation was viewed as a fundamental organismal change. With the aid of LGT, lineages are no longer restricted to exploring logically accessible niches; rather, they could acquire any set of genes – from any other organism, living at any place or time – to thrust them into completely novel, previously unavailable ecological contexts. The LGT effectively removes the barriers – both genetic and metaphorical – between bacterial taxa, allowing heritable information as well as evolutionary and ecological potential to flow between them. It is this seemingly unbounded potential for taxonomic mixis that can be puzzling. If LGT is the norm for genome evolution, why isn't the world filled with hopeful chimeras? Why do organisms fall into groups with shared properties? Shouldn't a group of chimeras, descended from a long line of chimeras, be impossible to classify in this way? Is LGT a convenient excuse for unexplainable genomic phenomena? Such wariness is justified, but progress over the past decade has shown that while gene transfer can be a powerful source of genetic innovation, it does not necessarily lead to phylogenetic chaos (Gogarten *et al.*, 2002; Lawrence and Hendrickson, 2003). For example, there are constraints on which genes are transferred with ease and which are more recalcitrant, as encapsulated by the complexity hypothesis (Jain *et al.*, 1999), or mobile genome hypothesis leading to a broad pan-genome (Tettelin *et al.*, 2005). In addition, there are apparent

constraints on the phylogenetic breadth of donors and recipients (Beiko *et al.*, 2005), a conclusion validated by molecular mechanisms which may constrain gene transfer (Hendrickson and Lawrence, 2006; Lawrence and Hendrickson, 2003, 2004). Both sets of constraints have served to maintain order among bacterial genomes in the face of gene transfer. That is, apparent phylogenetic order (Brown *et al.*, 2001; Fitz-Gibbon and House, 1999; Snel *et al.*, 1999; Tekaia *et al.*, 1999) does not lead to the conclusion that the impacts and lateral gene transfer have been overstated; rather, there are rules governing gene transfer that have served to preserve the phylogenetic hierarchy we observe.

1.2. GENOME EVOLUTION IN PATHOGENS

The first views of genome evolution arose from traditional models of gene evolution; that is, genome evolution could be viewed as collective gene evolution. Population genetic studies of gene evolution have focused on the fate of mutations, which represent gradual changes in genes. The similarity in gene order among closely related bacteria seemingly reinforced the view that genomes are relatively stable and change content slowly. Yet the availability of complete genome sequences shows that gene content can vary widely among even very closely related strains (Konstantinidis and Tiedje, 2005). The framework of shared genes necessarily minimizes the apparent roles of gene gain, gene loss, genome rearrangement, and allelic replacement of genes via recombination with conspecific strains. More importantly, these are inter-related processes: examination of available genome sequences suggests that LGT has had more impact on genome evolution than simply expanding gene inventory. The tempo and targets of other processes – including rearrangement, gene loss, and gene replacement – are affected as well. In effect, gene acquisitions catalyze genome evolution by a wide array of mechanisms. Herein we discuss how genomes have become more fluid in their content and organization as a result of lateral gene transfer and associated processes.

1.3. DIRECT IMPACT OF GENE ACQUISITION

The most obvious impact of LGT is the introduction of novel genes into a genome. Because the acquisition of novel traits via gene acquisition differs markedly from phenotypic evolution via the modification of genes within the chromosome, the phenotypic, physiological, and evolutionary effects of

such transfers cannot be overstated. First, acquired genes will have had their functions honed by selection in their donor genome and can produce fully functional products immediately after arrival, thereby allowing effective competition in new niches. In contrast, the modification of existing genes can explore functions that are available within small numbers of mutational steps from the existing sequences. This is because, in population genetic terms, the genes do not provide robust functions during transition from one adaptive peak to another, making exploration of the adaptive landscape problematic. In addition, only those proteins available within the genome can be modified in this way, whereas LGT can introduce new protein families. Beyond the introduction of novel genes, gene transfer can introduce genes for complex functions, because more than one gene may be mobilized at the same time. These large regions of acquired DNA are termed pathogenicity islands or PAI and contained dozens of genes (Blum *et al.*, 1994; Hacker *et al.*, 1997; Hacker and Kaper, 2000). It is not feasible for this larger number of genes to have evolved by mutational processes simultaneously, even if paralogs were available in the genome as starting substrates. Importantly, the acquisition of large fragments of DNA is not confined to pathogens. Similar large fragments in other genomes have been termed genomic islands, or GEI, reflecting the ability of introgressed DNA to alter any organism's physiological toolbox (Dobrindt *et al.*, 2004). Functions required large numbers of genes – such as the synthesis of complex coenzymes (Lawrence and Roth, 1996a), or the suite of genes required for methanogenesis (Chistoserdova *et al.*, 1998) – can be introduced at one time. We will not dwell on the numerous functions encoded by PAI or GEI, as there are chapters in the volume discussing them in detail, but pathogenicity islands can contain encode enormously intricate functions, such as the synthesis and deployment of type III secretion systems (Ochman *et al.*, 1996). Dramatic changes may occur very quickly; for example, a pathogenic fungus is believed to have arisen in within the past few decades following the acquisition of pathogenicity islands (Friesen *et al.*, 2006). These adaptations may use functions that were refined elsewhere for different physiological functions. For example, the acquisition of siderophores allows pathogens to obtain critical ions present at low concentration (e.g., Fe^{2+} or HPO_4^{2-}); yet these functions evolved in non-pathogenic bacteria living in non-host-associated, ion-poor environments. The ability of LGT to introduce all of the genes required to perform complex functions provides the cell with a route for ecological adaptation that is, for all intents and purposes, entirely unavailable via traditional gene modification by mutation and selection. This forces microbiologists to consider avenues of phenotypic evolution once thought to be unconventional.

1.4. THE ROLE OF GATEWAY GENES

While the direct effects of gene gain can be extensive, the real potential for genome change lies in the long-term, indirect effects. The addition of new genes to a bacterial genome results in numerous related, downstream events that may change the evolutionary trajectory of bacterial chromosomes as much as the gene acquisition event itself. In the broadest sense, gene gain can cause a dramatic change in the ecological niche exploited by the cell. Consider that the foreign genes we observe in bacterial chromosomes represent only a small fraction of those that were introduced into that cell's cytoplasm and recombined into a stable replicon. After this occurs, all gene acquisition events are filtered, so to speak, by natural selection. Genes which provide no useful function, or those that are actively problematic, will not be retained. Others may persist for short periods of time when their benefits are transient; for example, we see such evolutionary lability with antibiotic resistance genes in most lineages, which can be lost as readily as they are gained (Kirkup and Riley, 2004).

In very few cases, incoming genes will provide a function that is beneficial to the cell and can be retained for long periods of time (Lawrence and Ochman, 1997, 1998). In this context, the new gene inventory will predispose a cell for a substantially different lifestyle than that of its ancestor. As a result, the suite of incoming genes that would be potentially beneficial to the cell will also differ. That is, each gene acquisition event alters the potential fitness contributions of other incoming genes, effectively making genomes 'moving targets' for LGT. Genes which would otherwise have provided no benefit to the cell may become highly advantageous once that cell has acquired other, niche-altering genes. When acquired genes change the adaptive landscape to such a large degree, we can consider them to be 'gateway' genes, differentially conditioning the cell to the acquisition of other genes. Many consider initially acquired pathogenicity islands to be gateway genes, transforming an otherwise commensal organism into a pathogen (although see below for further discussion of this point). The adoption of the pathogenic lifestyle then favorably disposes the cell to the acquisition of other pathogenicity islands. For example, genomic analysis of the plant pathogen *Erwinia carotovora* suggests that they acquired genes required for life in and around plants; these genes include those involved in type III protein secretion, phytotoxin production, plant-cell adhesion, and nitrogen fixation (Toth *et al.*, 2006). These functions complement other acquired genes – such as those allowing for degradation of plant cell-wall polysaccharides, or the production of siderophores – in the making, a formidable pathogen, but one attacking plants, not animals. The

route towards plant pathogenesis was imparted by acquisition of a gateway gene cluster opening up this ecological niche to the nascent *Erwinia* lineage.

1.4.1. Providing Sites for Additional Gene Insertion

The integration of large regions of DNA can facilitate additional gene acquisition events in a passive way by providing non-disruptive sites wherein incoming genes can insert. Bacterial chromosomes are information rich, and little intergenic space is available for the insertion of new DNA without disrupting the expression of existing genes. Some bacteriophages circumvent this constraint by using integrases that target tRNA genes, and recombination there is non-disruptive, because phage-borne sequences reconstruct the tRNA, thereby masking the effects of phage insertion. Lacking this gene-specific option, random insertion of DNA will have minimal impact when the host genes that are disrupted upon integration in to the chromosome are of little or no value. Recently acquired DNA contains a high fraction of genes that do not persist for long periods of time (Lawrence and Ochman, 1998), making them conveniently dispensable targets. Over time, less useful genes will be culled by deletion, but shortly following acquisition they likely provide the bulk of the available sites for further insertions. As a result, higher rates of foreign gene integration can facilitate further gene acquisition, a relationship resembling a positive feedback loop.

1.4.2. Promotion of Genome Rearrangement

Just as newly acquired genes may allow for additional insertions, these sequences may also promote intragenomic rearrangements such as translocations and inversion. Just as they may act as neutral sites for insertion, they may act as places where inversion and translocations will be minimally disruptive. But beyond acting passively, newly acquired DNA may play a more active role in genomic rearrangements in two ways. First, they may provide catalysts in the form of integrases and transposases whose expression leads to genomic rearrangement. Second, they may provide more active substrates for recombination in the form of repeated DNA sequences. The genomes of closely related species of *Bordetella* illustrate the potential for genomic rearrangement (Parkhill *et al.*, 2003). *B. pertussis* and *B. parapertussis* are the causative agents of whooping cough, while *B. bronchiseptica* is implicated in more benign bronchial infections. Therefore all three organisms are pathogens, although *B. pertussis* has increased its virulence. Whereas the genomes of *B. bronchiseptica* and *B. parapertussis* are largely syntenic,

the *B. pertussis* genome shows a remarkable degree of rearrangement relative to its sibling species (Parkhill *et al.*, 2003). In addition, the *B. pertussis* genome has a large number of transposable elements; these IS elements may have facilitated the formation of inversions and translocations either directly, via transposition, or indirectly by providing sites of action for homologous recombination. While other factors – such as decreases in population size – may have contributed to the inability of *B. pertussis* to counter-select the accumulation of transposons in its genome, it is clear that the presence of IS elements can have a large impact on genome structure.

1.4.3. Effecting Recombination Interference

Following their introduction, laterally transferred genes will begin to evolve according to the directional mutations pressures of their recipient genome (Lobry and Sueoka, 2002; Sueoka, 1988, 1992). But foreign genes and native genes do not evolve independently. Closely related strains of bacteria can exchange genes by homologous recombination, a process termed allelic replacement. Here, sequences that retain a high degree of similarity may recombine if mechanisms of gene exchange – transduction, conjugation, or transformation – introduces the DNA into the cytoplasm of a conspecific strain. Such recombination can be beneficial within bacterial species by allowing the rapid dissemination of advantageous mutations between strains, or in allowing for the repair of deleterious mutations using functional genes from conspecific strains. The impact of recombination is not limited to simple gene conversion; depending on the physical limitations of the mechanisms of gene transfer – for example, the size of transducing particles – strains may also exchange small insertions, rearrangements or deletions. Therefore, inter-strain recombination is an active force in shaping genome composition. Despite acting on shared sequences, this process is not unaffected by acquisition of foreign DNA. As discussed above, gene acquisition will lead to a change in the organism's ecology, so that the newly altered recipient cell no longer occupies the same niche as its ancestor. Yet homologous recombination acts among all organisms whose DNA is sufficiently similar to allow for strand invasion, regardless of their underlying ecology. As populations of bacteria acquire different laterally transferred genes, they will begin to exploit sets of different niches, all the while sharing numerous genes that are not performing any niche-specific functions. Yet recombination at these shared loci will be suppressed if these recombination events affect neighboring, adaptive loci that enable the different lineages to exploit different niches. That is, some recombination events will result in poorly adapted hybrids

and will be counter-selected; the elimination of these less-fit cells from the population leads to lower rates of recombination being observed at loci flanking adaptive genes (Lawrence, 2002). This recombination interference does not affect all of the genes that are shared between newly diverging lineages; recombination will occur normally at shared loci that are not linked to sites of adaptive differences between strains because these recombinants will not suffer any fitness costs. In effect, the acquisition of foreign DNA will act to promote genetic isolation at their flanking loci, those genes which are shared between ecologically distinct strains. Eventually, complete genetic isolation will occur when there is no part of the chromosome that is not linked to an adaptive locus. As recombination rates decrease, point mutations will accumulate which impose pre-mating isolation via mismatch correction systems (Majewski and Cohan, 1999, 1998; Vulic *et al.*, 1999). One can view gene acquisition as an arbiter of bacterial speciation by catalyzing both genetic isolation and the accumulation of mutations in orthologous genes.

1.4.4. Promotion of Gene Loss

Not surprisingly – because organismal genomes are not constantly increasing in size – the gain of genes through horizontal gene transfer leads to gene loss. This occurs for three reasons. First, gene losses can be beneficial, if expression of those genes conflicts with the cell's new lifestyle. For example, losses of the *Shigella cadA* and *ompT* genes were beneficial in that expression of either gene decreases pathogenicity (Day *et al.*, 2001; Nakata *et al.*, 1993). This sort of gene loss is an inevitable consequence of the cell's changing ecological role. Second, gene loss can be neutral. The process of niche reorganization described above will render unimportant those genes which do not contribute to the cell's new lifestyle; without selection for function they will accumulate mutations and will eventually be deleted from the genome. There appears to be a strong "deletion" bias in bacterial genomes that prevents the accumulation of transposons, pseudogenes, and other segments of DNA that are not under positive selection for function (Lawrence, 2001; Mira *et al.*, 2001). As a result, bacterial genomes are information rich, with little intergenic spaces. Therefore, it is not surprising that, once selection is relaxed on some genes following that organism's exploitation of a new lifestyle, genes will be lost by deletion rather than persisting in the genome as pseudogenes. This process does not involve a targeting of specific genes for deletion. Rather, deletions that would have removed these genes from ancestral populations would have been counter-selected, preventing deleted strains from rising to high frequency. But following LGT, some deletions will no longer be detrimental, and genes will be lost. Pseudogenes appear in

high frequency only in those genomes where gene deletion has not kept pace with the rapid rate at which genes became useless, such as in the genomes of the obligate pathogen *Mycobacterium leprae* (Cole *et al.*, 2001). Last, gene loss may be detrimental, but insufficiently so to be prevented; from a population genetics standpoint, gene gain will, on average, necessitate gene loss. Beyond the loss of non-selected DNA, the cell cannot retain additional information in the form of newly acquired genes without sacrificing its ability to retain information elsewhere (Lawrence *et al.*, 1999; Lawrence, 2001). Put simply, a population of organisms can maintain only a finite amount of information under selection at any one time. When mutations in some genes are counter-selected (that is, organisms bearing mutations in these genes are eliminated from the population), then mutations must accumulate in other genes. Those mutations with the most detrimental effects will be eliminated and, by definition, effectively neutral mutations will be allowed to accumulate in their stead. This loss of information being maintained by selection can be manifested by the loss of entire genes. From a population genetic standpoint, although the functions some genes provide may be beneficial to a small degree, they will be insufficiently beneficial to prevent their loss.

1.5. IMPACT OF POPULATION SIZE

The limitations on information content are imposed by the population structure of the bacterium itself (Lawrence *et al.*, 1999; Lawrence, 2001). The exploitation of pathogenic lifestyles is often associated with a decrease in population size, with a commensurate decrease in information content. Bacteria with large population sizes and high rates of recombination can maintain large amounts of information; this large information content may be manifested in large numbers of genes. In contrast, organisms with small population sizes and small recombination rates can maintain far less information. Typically, these organisms have fewer genes as a result. As organisms transit between large and small effective population sizes, genome reduction will occur whereby the majority of genes in their genome may be lost. As expected, extreme genome reduction is seen in pathogens and symbionts with small population sizes, including *Mycoplasma* and *Buchnera* (Andersson, 2000; Andersson and Andersson, 1999b, 1999a; Razin, 1997). Here, extreme ecological specialization has led to very small effective population sizes and low rates of recombination. Because deleterious mutations are less effectively removed from the population, the numbers of genes that can be retained in the face of Muller's ratchet has decreased. Combined with the ever-present process of gene deletion, the genomes have been reduced to very small sizes. In one case, an insect symbiont has been reduced to

~160 kilobases in size (Nakabachi *et al.*, 2006). Aside from these extreme examples, it is often the case the adoption of the pathogenic lifestyle promoted by gene acquisition is accompanied by a reduction in population size. As a result, information content is limited and genomes decrease in size, although to a less dramatic extent. Comparison of pathogenic bacteria with their commensal relatives often shows a decrease in genome size.

1.6. THE FATE OF GENOMIC ISLANDS

The impact of GEI goes beyond the introduction of novel genes. In considering successful islands – those which contain genes being retained under purifying selection – there are two fates, both of which affect genome structure. First, an island may be modified. An island may be formed after the lateral transfer of any segment of DNA into a recipient cell. Following transfer, only those genes under selection will be retained; other genes will be lost to deletion, leading to collapse of the genomic island into a more compact form, bearing only those genes which contribute to the function (or functions) under selection. Because the mechanisms of gene transfer are more effective and efficient in mobilizing smaller segments of DNA, the compact island will be more likely to move to another genome. That is, the size of the island benefits the constituent genes, a property that has been termed the "selfish operon" (Lawrence and Roth, 1996b; Lawrence, 1997, 1999, 2003). Island reorganization may include the inclusion of genes native to this new genome, thereby augmenting the physiological capabilities of the island. Alternatively, genomic islands may decay, both losing some genes and distributing other genes throughout the chromosome. Genomic islands contain clusters of adjacent genes and operons because they arrived that way, not necessarily because there is any advantage to the recipient cell for that organization to be retained; the selfish operon hypothesis asserts that this clustering is beneficial to the genes in allowing for greater mobility. The distribution of these genes throughout the genome will contribute to genome heterogeneity. It is not clear what fraction of laterally transferred genes currently observed in bacterial genomes once arrived in larger genomic islands.

1.7. IS CHROMOSOME ARCHITECTURE A CATALYST TO GENOME REARRANGEMENT?

Although the genes contained in genomic islands may disperse throughout the genome, why would they? What advantage could there be to

translocation, inversion, or gene loss (beyond the removal of a gene whose product interferes with the cell's new physiology)? More subtle impacts of gene acquisition are evident when one considers genomes as more than collections of genes. What has become increasingly clear is that bacterial chromosomes are well-organized molecules which carry information required for their effective replication and partitioning into daughter cells (Hendrickson and Lawrence, 2006). For example, the FtsK translocase transports DNA across the division septum by moving along DNA towards the replication terminus (Bigot *et al.*, 2005; Capiaux *et al.*, 2002; Levy *et al.*, 2005); to do so, it recognizes strand-biased oligomers – termed AIMS, or architecture-imparting sequences – that are preferentially found on leading strands, and increase in abundance towards the terminus (Hendrickson and Lawrence, 2006; Lawrence and Hendrickson, 2003). The identity of AIMS is lineage specific, and different bacterial divisions use different sequences to perform these functions (Hendrickson and Lawrence, 2006; Lawrence and Hendrickson, 2004). Because the directionality of FtsK is imparted by the distribution and abundance of AIMS within replicores, the addition of genomic islands can disrupt chromosome partitioning by introducing DNA (a) whose AIMS are primarily in the improper orientation, (b) whose AIMS are abundant on both the leading and lagging strands, or (c) which is depauperate in AIMS used by the recipient cell. Here, the potentially detrimental effects of gene gain – not problematic in terms of the gene products they encode, but simply from the standpoint that incoming DNA can disrupt an otherwise well-ordered replicore – can be diminished if genomic islands are broken up, and the genes distributed to multiple locations. Similarly, gene loss could be catalyzed because the removal of improperly oriented AIMS is advantageous, even if removing the gene product is not. Thus, chromosome rearrangements may be advantageous following LGT by minimizing the effect of AIMS disruption.

1.8. IS GENOME FLUIDITY IN PATHOGENS TYPICAL OF OTHER BACTERIA?

One may ask if the inferences we make about the evolution of pathogen genomes are extensible to other, non-pathogenic organisms. We believe that studying the genomes of pathogens can lead to broad lessons regarding the evolution of all bacteria for several reasons. First, the evolution of pathogens from non-pathogenic ancestors can be framed as an adaptation to a new niche, something that certainly typifies lineage diversification in non-pathogenic organisms. One may argue that the adoption of a pathogenic lifestyle is a more rapid or dramatic change than adaptation to other lifestyles,

but this stance assumes that bacteria invade only niches very similar to their own. The acquisition of genomic islands is seen in many organisms, and LGT itself is prevalent in the genomes of all organisms that are not obligate intracellular parasites or symbionts. In this way, the evolution of pathogenicity can be seen as a model for any ecological shift. Second, one may argue that pathogens appear to proceed through a well-ordered ecological succession that is not obvious in other organisms (Lawrence, 2005). That is, the cascade of events that follow the acquisition of pathogenicity islands and subsequent adaptations could be fashioned into a series of stages that delineate a possible history of life. Initial events of LGT introduce genes disposing the cell to a pathogenic lifestyle. Subsequent gene gains and losses may follow facilitating further specialization to particular hosts (e.g., *Mycobacterium tuberculosis* or *Bordetella bronchiseptica*). The accompanying decrease in population size leads to gene loss, pseudogene formation, accumulation of transposable elements and extensive genomic rearrangements (e.g., *M. leprae* or *B. pertussis*). The decrease in population size and increasing rarity of recombination precludes the retention of large numbers of genes and genome reduction takes place, leading to highly reduced genomes (e.g., *Mycoplasma*). Such a drastic change in genome architecture can be difficult to reverse. While one can observe clades of pathogens with taxa in these various states of genomic decay, this is not the case with most non-pathogenic bacteria. Two points here are salient. First, it is not clear how many lineages experience genome decay as described here, and whether the fraction of pathogen lineages experiencing it is any larger than the fraction of non-pathogen lineages. Second, it is not at all clear how to assort organisms into pathogen and non-pathogen lineages since, as described below, pathogenicity itself is a difficult lifestyle to define rigorously.

1.9. WHEN IS A PATHOGEN NOT A PATHOGEN?

Pathogens are not the only organisms that live in close contact with hosts and whose evolution is contoured by the necessity to continuously adapt to a host's system. Commensal bacteria – classically represented by *Escherichia coli* – live in close quarters with their mammalian hosts and typically do neither harm nor good. Symbiotic bacteria provide recognizable and tangible benefits to their eukaryotic hosts, as do members of the Rhizobiaceae to their legume hosts in the form of fixed nitrogen, or *Buchnera* to its aphid host in the form of essential amino acids. Is it clear which organisms are pathogens and which are not? While members of the order Rhizobiales are often thought of as symbionts, other members include important pathogens

Table 1.1. *Lifestyles of the members of the order Rhizobiales*

Family	Species	Commonly described lifestyle
Bartonellaceae	*Bartonella henselae*	Cause of cat scratch fever in humans
Bartonellaceae	*Bartonella bovis*	Mammalian pathogen common among cattle
Bradyrhizobiaceae	*Bradyrhizobium japonicum*	A nitrogen-fixing symbiont of soybean plants
Bradyrhizobiaceae	*Rhodopseudomonas palustris*	Free-living photosynthetic bacterium
Brucellaceae	*Brucella suis*	True zoonotic human intracellular pathogen
Brucellaceae	*Brucella abortus*	True zoonotic and bovine pathogen
Phyllobacteriaceae	*Mesorhizobium loti*	Nitrogen-fixing symbiont of legumes
Rhizobiaceae	*Agrobacterium tumefaciens*	Parasitic cause of crown galls in plants
Rhizobiaceae	*Sinorhizobium meliloti*	Nitrogen-fixing symbiont of legumes

Representative species are listed along with their commonly discussed lifestyles.

such as members of the Bartonellaceae and Brucellaceae (Table 1.1). What is common to nearly all of these lineages is an ability to live among, and adapt to, eukaryotic hosts. The appearance of pathogens, symbionts, and free-living lineages in the same clade speaks to a fluid passage between these alternate lifestyles. Genome analyses have blurred the lines between pathogen, commensal, and symbiont. The pathogen *Brucella suis* has retained the capacity to use plant-derived compounds, possibly to aid in survival outside of animal hosts (Paulsen *et al.*, 2002). The notorious human pathogen *Escherichia coli* O157:H7 – which carries large numbers of pathogenicity islands – is commensal in cows; this lifestyle difference can be attributed to differential gene expression (Rashid *et al.*, 2006). *Vibrio fischeri* has been characterized as an elegant symbiont with its *Euprymna scolopes* host (Visick and Ruby, 2006). The squid has a light organ colonized by the bacteria, which, when they reach a sufficient density, use a quorum-sensing mechanism to flood the organ with light. The squid then allows light to pass to the ocean below, thus eliminating its shadow and making it more difficult for predators to detect.

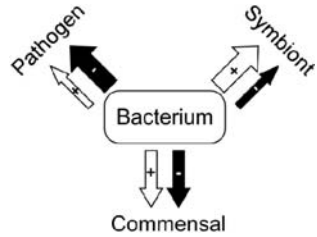


Figure 1.1. Interactions between a bacterium and eukaryotes. Open arrows denote beneficial interactions; dark arrows denote detrimental interactions.

In return, the squid provides the bacteria with food and shelter from predators. Yet luminescence has also been found in the cold-water fish pathogen *Vibrio salmonicida* (Fidopiastis *et al.*, 1999). In addition, the opportunistic fatal human pathogen *Vibrio vulnificus* – often acquired from eating raw oysters – and the Albensis serovar of *Vibrio cholerae* are also bioluminescent (Oliver *et al.*, 1986). These data suggest either that bioluminescence is important for pathogens, or that these pathogens are also symbionts in undiscovered contexts. Many symbionts and commensal bacteria rely upon the same adaptations to host defense systems that have previously been thought of as virulence genes (Hejnova *et al.*, 2005; Paulsen *et al.*, 2002). Commensal bacteria have been found to have both deleterious and advantageous effects on their hosts; for example, intestinal commensals lead to both proper development and to inflammatory bowel disease (Yan and Polk, 2004). In addition, commensal bacteria can provoke host reactions that are often attributed to pathogens defense (Zareie *et al.*, 2005). Symbiotic bacteria, primarily thought to be strictly beneficial, must ward off attack by the defense systems of their partner/hosts (Visick and Ruby, 2006). Genomic islands, antibiotic resistance elements, and other extragenomic elements have been implicated in rapid adaptation to novel pathogenic roles (Chapman *et al.*, 2006; Skyberg *et al.*, 2006), but it is not clear whether the significance we attribute reflects selection on those traits. From the perspective of the role of LGT, two points are important to make. First, the impact of a bacterium on a host organism represents the sum of all interactions (Figure 1.1). Have we misidentified the genes that are required to be a pathogen or have we oversimplified the complexity of being a bacterium? In some families of organisms there appear to be many pathogens closely related to commensals and symbionts, as seen in Table 1.1. Taken with the observation that some of the genes that we have characterized as being virulence genes are in found in non-pathogenic organisms, one might wonder what it means to term a particular organism a

commensal, pathogen, or symbiont. Far from a semantic argument, understanding the ecology of bacteria is a necessary precursor to understanding the significance of both laterally acquired genes and the subsequent genome changes that ensue.

1.10. ONE CELL'S PATHOGEN IN ANOTHER'S COMMENSAL

More to the point, because the net benefit or detriment of a host-microbe interaction depends on the host, a microorganism can act as a pathogen in one situation, as a symbiont in another, and as a harmless commensal in a third (Figure 1.1). For example, because we tend to observe the hosts and interactions we perceived as the most valuable, our assessment is certainly not unbiased. In the end, a simple lack of information may preclude easy delineations of organisms as pathogens or symbionts. We are largely blind to the myriad of signals, chemical or otherwise, that are exchanged between host and microbe, understanding them in insufficient detail to make qualified judgments about the nature of the relationship between them. In addition, relationships between a pair of organisms may be complicated by changes in their shared history. For example, pathogenic lineages of *Vibrio* and *Agrobacterium* likely arose from symbionts or commensals, and the reverse may also be true in other taxa. Last, host microbe interactions are difficult to parse meaningfully because the organisms themselves are still playing at the same game. Hiding one's intentions for as long as possible (from the bacterium's perspective) may buy the time needed to reproduce and obtain the necessary quorum to act more aggressively. Refinement of this interaction may lead to the development of more commensal or even symbiotic relationships. Far from being intractable, there is promise that genomic analyses will shed light on the interactions between bacteria and their hosts. Microarray studies can provide insight into gene expression patterns of both partners in these relationships and hold the promise of revealing responses in the chemical dialogue between hosts and their pathogenic, symbiotic, and commensal bacteria. While there has been some initial difficulty in isolating mRNA specifically from the bacterium (Schoolnik, 2002), progress is being made. For example, interactions between mouse and the malaria parasite in vivo show very different responses from both sides depending on which mouse tissues are examined (Lovegrove *et al.*, 2006). In addition, expression changes in response to mouse cecal environments were observed in the commensal and probiotic bacteria *Bacteroides thetaiotaomicron* and *Bifidobacterium longum* (Sonnenburg *et al.*, 2006). Host cells induced the immune system's interferon-responsive genes when they were exposed to both

B. thetaiotaomicron and *B. longum*, suggesting that the host were responding to these bacteria as if they were not entirely beneficial. In response, *B. longum* appears to regulate host antibacterial proteins upon induction of host cytokine-responsive genes. These studies will begin to dissect the intimate dialogue taking place between hosts and their colonizers, enabling a more sophisticated understanding of host-microbe interactions and the roles laterally acquired genes play.

1.11. CONCLUSIONS

The acquisition of laterally transferred DNA has impact on chromosome content and organization beyond the mere addition of new genes. Additional chromosome modifications may result from the addition of repetitive DNA or enzymes responsible for DNA rearrangement, from fundamental changes in the bacterium's ecological niche and possible changes in population structure, from the new suite of genes whose gain or loss may now be beneficial, and from alterations in the underlying chromosome architecture. Current retrospective analyses of bacterial genome evolution are biased to a large degree by our perception of what roles laterally transferred gene play. But the evolutionary path an organism may take can be difficult or impossible to predict because the roles played by laterally transferred genes in bacterial lifestyles can be ambiguous. In the end, we may only be able to conclude that life is complex, and describing the processes that shape genome evolution from the patterns we observe requires a degree of sophistication we currently do not possess.

REFERENCES

Andersson, J. O. (2000). Evolutionary genomics: Is *Buchnera* a bacterium or an organelle? *Curr Biol*, **10**, R866–8.

Andersson, J. O., and Andersson, S. G. (1999a). Genome degradation is an ongoing process in *Rickettsia*. *Mol Biol Evol*, **16**, 1178–91.

Andersson, J. O., and Andersson, S. G. (1999b). Insights into the evolutionary process of genome degradation. *Curr Opin Genet Dev*, **9**, 664–71.

Beiko, R. G., Harlow, T. J., and Ragan, M. A. (2005). Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci USA*, **102**, 14332–7.

Bigot, S., Saleh, O. A., Lesterlin, C., *et al.* (2005). KOPS: DNA motifs that control *E. coli* chromosome segregation by orienting the FtsK translocase. *EMBO J*, **24**, 3770–80.

Blum, G., Ott, M., Lischewski, A., *et al.* (1994). Excision of large DNA regions termed pathogenicity islands from tRNA-specific loci in the chromosome of an *Escherichia coli* wild-type pathogen. *Infect Immun*, **62**, 606–14.

Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E., and Stanhope, M. J. (2001). Universal trees based on large combined protein sequence data sets. *Nat Genet*, **28**, 281–5.

Capiaux, H., Lesterlin, C., Perals, K., Louarn, J. M., and Cornet, F. (2002). A dual role for the FtsK protein in *Escherichia coli* chromosome segregation. *EMBO Rep*, **3**, 532–6.

Chapman, T. A., Wu, X. Y., Barchia, I., *et al.* (2006). Comparison of virulence gene profiles of *Escherichia coli* strains isolated from healthy and diarrheic swine. *Appl Environ Microbiol*, **72**, 4782–95.

Chistoserdova, L., Vorholt, J. A., Thauer, R. K., and Lidstrom, M. E. (1998). C1 transfer enzymes and coenzymes linking methylotrophic bacteria and methanogenic Archaea. *Science*, **281**, 99–102.

Cole, S. T., Eiglmeier, K., Parkhill, J., *et al.* (2001). Massive gene decay in the leprosy bacillus. *Nature*, **409**, 1007–11.

Day, W. A., Jr., Fernandez, R. E., and Maurelli, A. T. (2001). Pathoadaptive mutations that enhance virulence: genetic organization of the *cadA* regions of *Shigella* spp. *Infect Immun*, **69**, 7471–80.

Dobrindt, U., Hochhut, B., Hentschel, U., and Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol*, **2**, 414–24.

Doolittle, W. F. (1999a). Lateral genomics. *Trends Cell Biol*, **9**, M5–8.

Doolittle, W. F. (1999b). Phylogenetic classification and the universal tree. *Science*, **284**, 2124–9.

Fidopiastis, P. M., Sorum, H., and Ruby, E. G. (1999). Cryptic luminescence in the cold-water fish pathogen *Vibrio salmonicida*. *Arch Microbiol*, **171**, 205–9.

Fitz-Gibbon, S. T., and House, C. H. (1999). Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res*, **27**, 4218–22.

Friesen, T. L., Stukenbrock, E. H., Liu, Z., *et al.* (2006). Emergence of a new disease as a result of interspecific virulence gene transfer. *Nat Genet*, **38**, 953–6.

Gogarten, J. P., Doolittle, W. F., and Lawrence, J. G. (2002). Prokaryotic evolution in light of gene transfer. *Mol Biol Evol*, **19**, 2226–38.

Goldschmidt, R. (1940). *The material basis of evolution*. New Haven, CT: Yale University Press.

Groisman, E. A., and Ochman, H. (1994). How to become a pathogen. *Trends Microbiol*, **2**, 289–94.

Groisman, E. A., and Ochman, H. (1997). How *Salmonella* became a pathogen. *Trends Microbiol*, **5**, 343–9.

Hacker, J., and Kaper, J. B. (2000). Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol*, **54**, 641–79.

Hacker, J., Blum-Oehler, G., Mühldorfer, I., and Tschäpe, H. (1997). Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol Microbiol*, **23**, 1089–97.

Hejnova, J., Dobrindt, U., Nemcova, R., *et al.* (2005). Characterization of the flexible genome complement of the commensal *Escherichia coli* strain A0 34/86 (O83 : K24 : H31). *Microbiology*, **151**, 385–98.

Hendrickson, H., and Lawrence, J. G. (2006). Selection for chromosome architecture in bacteria. *J Mol Evol*, **62**, 615–29.

Jain, R., Rivera, M. C., and Lake, J. A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci USA*, **96**, 3801–6.

Kirkup, B. C., and Riley, M. A. (2004). Antibiotic-mediated antagonism leads to a bacterial game of rock-paper-scissors in vivo. *Nature*, **428**, 412–4.

Konstantinidis, K. T., and Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA*, **102**, 2567–72.

Kurland, C. G. (2000). Something for everyone. Horizontal gene transfer in evolution. *EMBO Rep*, **1**, 92–5.

Lawrence, J. G. (1997). Selfish operons and speciation by gene transfer. *Trends Microbiol*, **5**, 355–9.

Lawrence, J. G. (1999). Selfish operons: the evolutionary impact of gene clustering in the prokaryotes and eukaryotes. *Curr Opin Genet Dev*, **9**, 642–8.

Lawrence, J. G. (2001). Catalyzing bacterial speciation: correlating lateral transfer with genetic headroom. *Syst Biol*, **50**, 479–96.

Lawrence, J. G. (2002). Gene transfer in bacteria: speciation without species? *Theor Popul Biol*, **61**, 449–60.

Lawrence, J. G. (2003). Gene organization: selection, selfishness, and serendipity. *Annu Rev Microbiol*, **57**, 419–40.

Lawrence, J. G. (2005). Horizontal and vertical gene transfer: the life history of pathogens. *Contrib Microbiol*, **12**, 255–71.

Lawrence, J. G., and Hendrickson, H. (2003). Lateral gene transfer: when will adolescence end? *Mol Microbiol*, **50**, 739–49.

Lawrence, J. G., and Hendrickson, H. (2004). Chromosome structure and constraints on lateral gene transfer. *Dyn Genet*, **2004**, 319–36.

Lawrence, J. G., and Ochman, H. (1997). Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol*, **44**, 383–97.

Lawrence, J. G., and Ochman, H. (1998). Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci USA*, **95**, 9413–7.

Lawrence, J. G., and Roth, J. R. (1996a). Evolution of coenzyme B_{12} synthesis among enteric bacteria: evidence for loss and reacquisition of a multigene complex. *Genetics*, **142**, 11–24.

Lawrence, J. G., and Roth, J. R. (1996b). Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics*, **143**, 1843–60.

Lawrence, J. G., Roth, J. R., and Charlebois, R. L. (1999). *Genomic flux: genome evolution by gene loss and acquisition*. Washington, DC: American Society for Microbiology.

Levy, O., Ptacin, J. L., Pease, P. J., *et al.* (2005). Identification of oligonucleotide sequences that direct the movement of the *Escherichia coli* FtsK translocase. *Proc Natl Acad Sci USA*, **102**, 17618–23.

Lobry, J. R., and Sueoka, N. (2002). Asymmetric directional mutation pressures in bacteria. *Genome Biol*, **3**, RESEARCH0058.

Lovegrove, F. E., Pena-Castillo, L., Mohammad, N., *et al.* (2006). Simultaneous host and parasite expression profiling identifies tissue-specific transcriptional programs associated with susceptibility or resistance to experimental cerebral malaria. *BMC Genomics*, **7**, 295.

Majewski, J., and Cohan, F. M. (1998). The effect of mismatch repair and heteroduplex formation on sexual isolation in *Bacillus*. *Genetics*, **148**, 13–8.

Majewski, J., and Cohan, F. M. (1999). DNA sequence similarity requirements for interspecific recombination in *Bacillus*. *Genetics*, **153**, 1525–33.

Mira, A., Ochman, H., and Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends Genet*, **17**, 589–96.

Nakabachi, A., Yamashita, A., Toh, H., *et al.* (2006). The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science*, **314**, 267.

Nakata, N., Tobe, T., Fukuda, I., *et al.* (1993). The absence of a surface protease, OmpT, determines the intercellular spreading ability of *Shigella*: the relationship between the *ompT* and *kcpA* loci. *Mol Microbiol*, **9**, 459–68.

Ochman, H., Soncini, F. C., Solomon, F., and Groisman, E. A. (1996). Identification of a pathogenicity island required for *Salmonella* survival in host cells. *Proc Natl Acad Sci USA*, **93**, 7800–4.

Oliver, J. D., Roberts, D. M., White, V. K., Dry, M. A., and Simpson, L. M. (1986). Bioluminescence in a strain of the human pathogenic bacterium *Vibrio vulnificus*. *Appl Environ Microbiol*, **52**, 1209–11.

Parkhill, J., Sebaihia, M., Preston, A., *et al.* (2003). Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nat Genet*, **35**, 32–40.

Paulsen, I. T., Seshadri, R., Nelson, K. E., *et al.* (2002). The *Brucella suis* genome reveals fundamental similarities between animal and plant pathogens and symbionts. *Proc Natl Acad Sci USA*, **99**, 13148–53.

Rashid, R. A., Tabata, T. A., Oatley, M. J., *et al*. (2006). Expression of putative virulence factors of *Escherichia coli* O157:H7 differs in bovine and human infections. *Infect Immun*, **74**, 4142–8.

Razin, S. (1997). The minimal cellular genome of *Mycoplasma*. *Indian J Biochem Biophys*, **34**, 124–30.

Schoolnik, G. K. (2002). Functional and comparative genomics of pathogenic bacteria. *Curr Opin Microbiol*, **5**, 20–6.

Skyberg, J. A., Johnson, T. J., Johnson, J. R., *et al*. (2006). Acquisition of avian pathogenic *Escherichia coli* plasmids by a commensal *E. coli* isolate enhances its abilities to kill chicken embryos, grow in human urine, and colonize the murine kidney. *Infect Immun*, **74**, 6287–92.

Snel, B., Bork, P., and Huynen, M. A. (1999). Genome phylogeny based on gene content. *Nat Genet*, **21**, 108–10.

Sonnenburg, J. L., Chen, C. T., and Gordon, J. I. (2006). Genomic and metabolic studies of the impact of probiotics on a model gut symbiont and host. *PLoS Biol*, **4**, e413.

Sueoka, N. (1988). Directional mutation pressure and neutral molecular evolution. *Proc Natl Acad Sci USA*, **85**, 2653–7.

Sueoka, N. (1992). Directional mutation pressure, selective constraints, and genetic equilibria. *J Mol Evol*, **34**, 95–114.

Tekaia, F., Lazcano, A., and Dujon, B. (1999). The genomic tree as revealed from whole proteome comparisons. *Genome Res*, **9**, 550–7.

Tettelin, H., Masignani, V., Cieslewicz, M. J., *et al*. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci USA*, **102**, 13950–5.

Toth, I. K., Pritchard, L., and Birch, P. R. (2006). Comparative genomics reveals what makes an enterobacterial plant pathogen. *Annu Rev Phytopathol*, **44**, 305–36.

Visick, K. L., and Ruby, E. G. (2006). *Vibrio fischeri* and its host: it takes two to tango. *Curr Opin Microbiol*, **9**, 632–8.

Vulic, M., Lenski, R. E., and Radman, M. (1999). Mutation, recombination, and incipient speciation of bacteria in the laboratory. *Proc Natl Acad Sci USA*, **96**, 7348–51.

Yan, F., and Polk, D. B. (2004). Commensal bacteria in the gut: learning who our friends are. *Curr Opin Gastroenterol*, **20**, 565–71.

Zareie, M., Riff, J., Donato, K., *et al*. (2005). Novel effects of the prototype translocating *Escherichia coli*, strain C25 on intestinal epithelial structure and barrier function. *Cell Microbiol*, **7**, 1782–97.

CHAPTER 2

Bacterial Recombination in vivo

Xavier Didelot and Daniel Falush

2.1. INTRODUCTION

In eukaryotes, the great majority of genetic recombination takes place during the complex and highly organized process of meiotic division, a part of sexual reproduction. As a consequence there are a number of constraints on patterns of variability in the recombination process. Recombination takes place only between organisms that are similar enough for their offspring to be viable, and therefore it is generally limited in the novelty it can introduce. Within species, the most common cause of reproductive isolation is geographical separation. In most higher animals and plants, the number of crossovers per chromosome is predictably a number between 1 and 5 in both sexes. Even where individuals differ in the amount of recombination that they initiate, for example because of the absence of crossing over in male *Drosophila*, the existence of a common mating pool will tend to homogenize the population with respect to the amount of genetic exchange that has occurred in the ancestry of each individual. In summary, while the mating process is elegant, eukaryotic recombination is typically quite predictable with minimal differences in genetic patterns between individuals in the same species.

In bacteria, there are no such rules. Recombination is never obligate and occurs by three distinct mechanisms; transformation, transduction, and conjugation, each of which in their nature can vary enormously between lineages. Recombination can involve import of DNA with high homology to existing sequence, but can also (sometimes simultaneously) introduce entirely novel DNA. Individual lineages do not necessarily contribute to, or draw their DNA from, a single mating pool, and therefore can diverge considerably in the composition of the DNA that they have acquired. Bacterial

recombination is thus variable and interesting. In this review we first describe features of the biology of exchange that promote differences in patterns and rates of recombination between different strains. Second, we review the evidence for gross differences in recombination rates between bacteria from different genera, as estimated using a variety of experimental approaches. Third, we describe a handful of examples which imply that patterns and rates of recombination can vary considerably between lineages in the same species. Finally, we discuss how we might go about acquiring a systematic knowledge of bacterial recombination patterns in nature.

2.2. BACTERIAL MECHANISMS OF EXCHANGE ALLOW ENORMOUS VARIABILITY IN RECOMBINATION RATES

Bacterial recombination is the process by which foreign DNA is inserted in the bacterial genome. This always happens in two steps: first a molecule of DNA is taken into the bacterial cell, and then this molecule (or some parts of it) is inserted in the genome.

Transformation is the process by which naked DNA from the external environment is incorporated into a bacterial cell. Successful transformation requires the recipient cell to exhibit on its membrane special DNA binding proteins, as well as various molecules that facilitate the incorporation of the DNA in the cell once it is bound to the membrane (Dubnau, 1999). Some species do not transform DNA at all. In many of the 44 bacterial species listed as transformable by Lorenz and Wackernagel (1994), only certain strains are and only at certain moments of their growth cycle. Moreover, certain bacterial species require that the DNA molecule contain specific short sequences for transformation to be efficient, sequences that are found with frequencies significantly higher than expected on their genomes (e.g., *Neisseria gonorrhoeae*, Goodman and Scocca, 1988).

Transduction involves the injection of a DNA molecule into the recipient cell by phage. Many phage infections will lead to the death of the bacterium and release of new phage into the environment. However, the phage capsid might contain DNA from the previous host instead of phage DNA, and this DNA can then recombine with bacterial chromosomal DNA (generalized transduction). Alternatively, the bacteriophage DNA may itself be inserted into the bacterial genome (specialized transduction). There are at least as many phages in the environment as there are bacteria (Chibani-Chennoufi *et al.*, 2004; Weinbauer, 2004), so in order to survive most bacteria need to be resistant to the great majority of phages that they encounter, but phages need to be able to infect at least one lineage to reproduce. These forces seem to have

led to rapid coevolution, making it possible to identify particular lineages of *Salmonella*, for example, according to their susceptibility to particular phages (Anderson and Williams, 1956; Le Minor, 1988). Transduction can only mediate genetic exchange between pairs of strains both of which are susceptible to the same phage. Transduction has been shown to happen in a large number of bacterial species. The vast majority of natural isolates of *Salmonella enterica* and *Escherichia coli* were shown to release at least one type of lysogenic phage, most of which were capable of transduction (Schicklmaier and Schmieger, 1995; Schicklmaier *et al.*, 1998). The frequency of lytic phage infection in vivo is intrinsically harder to investigate.

Sex is also associated with infection in conjugation, which happens when two bacterial cells come in contact, and a plasmid is transferred from one to the other. A plasmid is a genetic element that exists exclusively in a host cell and replicates independently from it. Plasmids are numerous and diverse (e.g., more than 300 different naturally occurring plasmids have been discovered from isolates of *Escherichia coli* alone) and can transmit to a wide variety of hosts (e.g., the Tn*916*-Tn*1545* plasmid family can propagate in 24 different bacterial genera, Clewell *et al.*, 1995). Davison (1999) cites more than 50 different studies as evidence that conjugation happens frequently in a variety of species and environmental conditions.

Once DNA has been incorporated into a recipient cell through transformation, transduction, or conjugation, it still needs to be inserted in the genome of the recipient cell for the recombination to be successful. This is usually done through homologous recombination, where the homology (at least in places) of the recombinant DNA and the host DNA favours base pairing. Homologous recombination is a complex mechanism, involving for example at least 25 different genes in *Escherichia coli*. The common denominator is the RecA protein coded by *recA* (Smith, 1988), which has been shown to be essential in nearly every recombination pathway and has been identified, in one form or another, in all bacterial species studied.

In principle, recombination of the bacterial genome can happen with any DNA molecule that has been taken up in the bacterial cell. However, most bacteria have a mismatch repair (MMR) system, which can prevent recombination of certain fragments. The MMR is usually encoded by the *mutS* and *mutL* genes and their homologues, which have been found in eukaryotes as well (Modrich and Lahue, 1996). One exception is *Helicobacter pylori*, which does not have a *mutL* gene (Tomb *et al.*, 1997) and whose *mutS* gene is not involved in mismatch repair (Bjorkholm *et al.*, 2001). In *Salmonella*, the mismatch repair system encoded by *mutS* and *mutL* is responsible for a 100- to 1000-fold decrease in the frequency of recombination between Typhi

and Typhimurium which are 98–99% identical at the nucleotide level (Zahrt and Maloy, 1997). On top of this, the *recD*-dependent exonuclease activity of RecBCD reduces this frequency by a similar rate (Zahrt and Maloy, 1997). Mutator strains, are found at significant frequency in natural populations of *Escherichia coli*. Many of these strains lack mismatch repair function (Tago *et al.*, 2005). In *Bacillus*, the mismatch repair system plays only a small role in the prevention of recombination compared to the mechanism of heteroduplex formation (Majewski and Cohan, 1998, 1999) and the homology constraints for fragments to be inserted are less stringent than in wild-type *Salmonella* strains. Whatever the mechanisms responsible for the prevention of recombination, the key determinant to the success of a recombination event seems to be the level of divergence between the imported DNA and the genomic DNA. Specifically, it was shown that the probability of success of a recombination event is inversely proportional to the exponential of the level of divergence in bacterial species as diverse as *Escherichia coli* (Vulic *et al.*, 1997), *Bacillus* (Majewski and Cohan, 1998), or *Streptococcus pneumoniae* (Majewski *et al.*, 2000).

2.3. QUALITATIVE DIFFERENCES IN LEVELS OF RECOMBINATION BETWEEN BACTERIAL GENERA

Until 15 years ago, even though the recombination mechanisms had been observed and analysed in laboratories, geneticists believed that bacterial recombination was rare in nature. This belief was due to the fact that only *Escherichia coli* had been extensively studied (cf. for example the first study of a bacterial population structure, Selander and Levin, 1980). We now know, however, that not all bacterial species have the same level of clonality as *Escherichia coli* and that bacterial population structures can be very diverse (e.g., Achtman, 2004).

It is the acquisition of multilocus enzyme electrophoresis (MLEE) data in several other species than *E. coli* that led to the discovery that not all bacterial species are principally clonal. Maynard Smith *et al.* (1993) applied an index of association denoted IA (first used by Brown *et al.*, 1980) to measure the level of correlation between loci throughout an MLEE dataset. I_A is equal to $V_O/V_E - 1$ where V_O is the observed variance for the distance between pairs of individuals and V_E is the expected variance for the distance between pairs of individuals when assuming that all loci are independent. If the loci are really independent (i.e., recombination is widespread so that there is no linkage disequilibrium), I_A will be close to 0, whereas if the population is mainly clonal, the index should show a strong association between

loci. Maynard Smith *et al.* (1993) found three different types of population structures. First, some species have a strong level of association even for subgroups, which indicates an essentially clonal history (e.g., *Salmonella*, which is a close relative to *E. coli*). Second, some species have a strong association index for the whole species but one close to 0 for subgroups, which indicates that recombination occurs between members of the same groups but not between groups (e.g., *Rhizobium*). Finally, in some species the level of association is low at all levels, which indicates a high level of recombination throughout the species, even between distant members of the species (e.g., *Neisseria* or *Helicobacter pylori*, Suerbaum *et al.*, 1998). Because the value of the index of association I_A does not depend only on the recombination rate (Maynard Smith *et al.*, 1993), they cannot be easily interpreted beyond this.

The availability of multilocus sequence typing (MLST, Maiden *et al.*, 1998) data from a wide range of species has stimulated development of new measures of recombination that are complementary and have clearer interpretations than the index of association described earlier. The first one, usually denoted r/m, is the ratio of probabilities that a given nucleotide be altered through either recombination or point mutation (first introduced by Guttman and Dykhuizen, 1994). It is therefore a measure of the effect of recombination relative to mutation, in terms of how many substitutions they cause. It is not to be confused with the second measure, denoted ρ/θ, which represents the ratio of frequency of recombination and mutation events (first introduced by Milkman and Bridges, 1990). If we consider each MLST gene fragment as either not having been affected by import or having been affected by a single import (which is acceptable since fragments are quite small, usually less than 500 bp), then ρ/θ is equivalent to the number of fragments altered through recombination relative to mutation, which explains why ρ/θ is sometimes called "the r/m per allele" as opposed to "the r/m per site" for what we defined as r/m (other names for ρ/θ include "c/μ" and "r/μ"). A few studies using either (or both) of these measures are summarized in Table 2.1, illustrating a wide diversity in the rate and effect of recombination between bacterial species.

2.3.1. Microevolutionary Methods

The most widely applied method of estimation of the relative importance of recombination and mutation is a counting method described by Feil *et al.* (2000). It requires an MLST dataset where the isolates are representative of the whole of the species. The data is first divided into clonal complexes and a member of each clonal complex is determined to be the most likely

Table 2.1. *Values of r/m and ρ/θ as calculated by different studies in a variety of bacterial species*

Species	r/m	ρ/θ	Reference
Escherichia coli		0.02	Milkman and Bridges, 1990
Escherichia coli	<50	0.08	Guttman and Dykhuizen, 1994
Neisseria meningitidis	100	3.6	Feil *et al.*, 1999
Streptococcus pneumoniae	45–53	10	Feil *et al.*, 2000
Neisseria meningitidis	275	16	Jolley *et al.*, 2000
Neisseria meningitidis	100	4.75	Feil *et al.*, 2001
Streptococcus pneumoniae	61	8.9	
Staphylococcus aureus	24	6.5	
Helicobacter pylori		>1	Falush *et al.*, 2001
Enterococcus faecium	24		Homan *et al.*, 2002
S. pneumoniae	24.2		Robinson *et al.*, 2002
Serogroup 6			
Campylobacter jejuni	47	8	Schouls *et al.*, 2003
Pseudomonas syringae		0.121–0.622	Sarkar *et al.*, 2004
Staphylococcus aureus		0.07	Feil *et al.*, 2003
Streptococcus pyogenes		1.4	McGregor *et al.*, 2004
Borrelia burgdorferi		2.7–3	Qiu *et al.*, 2004
Halorubrum	43	5	Papke *et al.*, 2004
Pseudomonas viridiflava	38		Goss *et al.*, 2005
Xylella fastidiosa	3.23	0.45	Scally *et al.*, 2005
Sulfolobus islandicus	6.8	1.4	Whitaker *et al.*, 2005
Neisseria meningitidis	6.2–16.8	0.16–1.83	Jolley *et al.*, 2005
Neisseria meningitidis		1.1	Fraser *et al.*, 2005
Streptococcus pneumoniae		2.1	
Staphylococcus aureus		0.11	
Campylobacter jejuni		0.3–1.2	Fearnhead *et al.*, 2005
Escherichia coli		0.3–2.1	Wirth *et al.*, 2006
Bacillus cereus		0.05	Hanage *et al.*, 2006
Helicobacter pylori		2.5	
Burkholderia pseudomallei		6.38	
Streptococcus pyogenes		7.21	
Neisseria meningitidis	5–8	0.7–1.2	Didelot and Falush, 2007
Bacillus	1.3–2.8	0.2–0.5	

founder, for example using the BURST algorithm (Feil *et al.*, 2004; Spratt *et al.*, 2004). The method then focuses on the single locus variants of robustly attributed clonal complex founders and tries to determine how they diverged from the founder. It is assumed that they diverged through mutation if the genetic difference is due to a single nucleotide and through recombination otherwise. Dividing the number of divergence events that take place through recombination and mutation respectively gives an estimator of ρ/θ while dividing the number of nucleotide differences introduced by recombination and mutation events is an estimator of r/m. An improvement of the method is to consider that if a single locus variant is different at just one nucleotide but the new allele is also found in some unrelated isolate of the sample, then the divergence was caused by recombination rather than mutation (Feil *et al.*, 2001). In any case, the reasonableness of these assumptions depends on the size of the MLST fragments, and the distribution of divergence levels within imported fragments. This method of estimation of r/m and ρ/θ focuses on the genetic variation found between close relatives within clonal complexes. It is therefore the relative contribution of recombination and mutation to the divergence of recent clones that is estimated. This method is reasonably robust to the definition of the population but produces good estimators of r/m and ρ/θ throughout the evolution of the species only if recombination and mutation played similar roles in recent and deep phylogeny.

A related method of estimation of the relative contribution of recombination and mutation was applied to *Helicobacter pylori* by Falush *et al.* (2001). It uses sequence data from pairs of strains isolated sequentially from the same host. This provides a method of identifying pairs of strains that are a priori likely to be closely related to each other. A close relationship is then confirmed or rejected by MLST. One advantage of this approach is that it works in the small number of bacterial species that evolve so rapidly that it is hard to find pairs of strains that are related at the allele level in the general population (Suerbaum *et al.*, 1998). Second, the method provides some information on the rate at which the bacteria evolve, especially if the biology of infection makes it likely that the mutation and recombination all took place in the same host and the time between isolation is a large fraction of the overall time of infection. Third, because the method does not require the use of most of the data just to identify candidate recombination or mutation events, it is easier to develop statistical methods that take full account of uncertainty in estimating parameters. However, for bacteria that evolve relatively slowly, a great deal of sequencing may be required in order to identify events. Specifically, Falush *et al.* (2001) designed a model for how bacterial strains evolve that uses three parameters: the rate of recombination, the rate

of mutation, and the average length of recombination fragments. The ratio of the first two quantities is a measure of ρ/θ, the relative rates of recombination and mutation that occurred during colonization. In this example, the average import fragment size was small enough that the three single nucleotide changes that were observed could potentially all have been caused by homologous recombination, showing that differentiating mutation and recombination can be hard in some species. This method showed a high level of recombination and also that the imported DNA was typical of DNA found in other *H. pylori* strains from the same location. Thus, there was little evidence of barriers to exchange within *H. pylori*, which is consistent with the apparent absence of mismatch repair function in the species (Bjorkholm *et al.*, 2001).

2.3.2. Population Genetics Methods

A different approach consists in considering a model of how the bacterial populations evolve under the inuence of recombination and mutation, and using all of the information in the dataset to jointly estimate the history of the sample and the parameters of the model. A difficulty is that the parameter space to explore is large. The evolution of DNA sequences in a sample from a homogeneous population evolving under the joint effects of mutation and intrapopulation recombination can be modelled using a process known as the ancestral recombination graph (ARG, first described in Hudson, 1983). Methods to perform full-likelihood inference under this model have been developed using importance sampling (Griffiths and Marjoram, 1996; Fearnhead and Donnelly, 2001) or Monte Carlo Markov chain (Kuhner *et al.*, 2000) but remain too slow to be applied on datasets from bacterial population genetics.

The model of recombination can be simplified by assuming that each import introduces new polymorphism at a constant rate to a single tract of sequence (Didelot and Falush, 2007). This approach makes inference of the clonal history of the species and the imports that have taken place tractable and provides reasonable estimates of the recombination rate (Didelot and Falush, 2007). One advantage of this method is that it does not assume that recombination happens exclusively between members of the modelled population.

Another approach to make to inference of the recombination rate possible using population genetic methods is to work on some summary statistics of the data rather than the whole data. An example of this is the method

used by Fraser *et al.* (2005) and Hanage *et al.* (2006), which summarises multilocus sequence datasets by the distribution of the number of allele differences between all pairs of strains. Using a simple model of bacterial population evolution through mutation and recombination, it is possible to write explicitly the likelihood of a distribution of allele distances given the mutation and recombination rates, and maximum likelihood methods can be used to infer the mutation and recombination rates from a dataset. This method is also robust to interpopulation imports but, because it works on allelic types instead of sequences and summarizes the data by a handful of numbers, may lose a large proportion of the information available in the data.

Another widely used summary statistics are the patterns of linkage disequilibrium (LD) in a dataset as was done for example by Jolley *et al.* (2005), Fearnhead *et al.* (2005), and Wirth *et al.* (2006). Every time a recombination event happens, it breaks down the correlation between the sites in and out of the recombined region changes. The idea is therefore to summarise a dataset using the levels of association between the nucleotide at all pairs of polymorphic sites. Statistical methods have been developed to estimate the rate $\rho/2$ at which recombination happened from looking at the patterns of LD (Hudson, 2001; McVean *et al.*, 2002). This estimate of the rate of recombination divided by an estimate of the rate of mutation $\theta/2$ (calculated for example using Watterson's moment estimator, Watterson, 1975) gives an estimate of ρ/θ. r/m can be estimated as $(\rho/\theta) \times t \times d$ where t is the average length of an import (Table 2.2 summarizes a few estimates of the average tract length) and t the average genetic distance of an import (i.e., the average pairwise distance of the species if we exclude interspecies imports and assume that recombination can happen with equal probabilities between any two members of the species).

2.3.3. Comparing Different Estimates

Great caution must be taken when comparing estimates of r/m and ρ/θ from different studies such as the ones shown in Table 2.1 because the methods used to estimate these quantities and described earlier have very different properties. Most efforts at estimating recombination rates to date have tried to estimate average rates across a single species. Because bacteria cannot be subdivided based on reproductive isolation in the way that eukaryotes are, the boundaries between species may not be clear, making the choice of appropriate samples for analysis difficult. The extent to which

Table 2.2. *A few estimates of the recombination fragment size (in base pairs)*

Species	Tract length	Reference
Escherichia coli	1000	Milkman and Bridges, 1990
Helicobacter pylori	259–732	Falush *et al.*, 2001
Neisseria meningitidis	1100	Jolley *et al.*, 2005
Campylobacter jejuni	225–750	Fearnhead *et al.*, 2005
Helicobacter pylori	100–325	
Bacillus spp.	193–435	Didelot and Falush, 2007

this is an issue varies between the methods. For example, if the isolates in a sample come from two distinct clades, A and B, then there is an issue as to whether recombination rates should be analyzed separately or together. For micro-evolutionary methods, which involve inspecting changes between closely related strains, the rate that will be obtained from analyzing the two clades together should be approximately the weighted average of the rates estimated considering them independently. Population-genetic methods might estimate a considerably higher rate in the joint analysis (if clades A and B are single lineages within a larger recombining population) or lower (if clades A and B are both freely recombining but mutually isolated gene pools).

The values obtained in population-genetic methods will only really be comparable to those obtained by micro-evolutionary approaches if they are applied to a sample from a population that has largely homogeneous rates of recombination within it and is largely isolated from external sources of DNA. Most studies do not attempt to identify the origin of imports or to decide whether the population under study is a single population and what fraction, if any of the imports come from extra-population sources. Some of the studies that have looked at interspecific recombination are shown in Table 2.3. For example, the high level of diversity of the *nanA* virulence gene in *Streptococcus pneumoniae* is in part due to recombination with members of another species, *S. oralis* (King *et al.*, 2005). Dingle *et al.* (2005) used the linkage model of STRUCTURE (Pritchard *et al.*, 2000; Falush *et al.*, 2003a) to detect both recent and ancient recombination events between *Campylobacter coli* and *C. jejuni* and found that they occurred at a low but significant rate. More rarely, recombination has also been observed between members of

Table 2.3. *Validity of assuming that the populations are clonal, closed, and unstructured in different bacterial species**

Species	Clonal	Closed	Unstructured	Reference
Neisseria meningitidis	–	–	–	Feil *et al.*, 1996
Mycobacterium tuberculosis	++	++	–	Kremer *et al.*, 1999
Chlamydia spp.	–	–	+	Millman *et al.*, 2001
Mycobacterium bovis	++	++	–	Smith *et al.*, 2003
Helicobacter pylori	–	++	–	Falush *et al.*, 2003b
Yersinia pestis	++	++	++	Achtman, 2004
Yersinia, other species	–	–	+	Kotetishvili *et al.*, 2005
Streptococcus pneumoniae	–	–	+	King *et al.*, 2005
Neisseria meningitidis	–	–	–	Hanage *et al.*, 2005
Campylobacter jejuni	–	+	+	Dingle *et al.*, 2005
Campylobacter coli	–	+	+	
Wolbachia supergroups	–	–	++	Baldo *et al.*, 2006
Salmonella enterica	–	+	++	Falush *et al.*, 2006
Escherichia coli	+	++	–	Wirth *et al.*, 2006
Bacillus spp.	+	–	+	Didelot and Falush, 2007

* ++, +, and − indicate respectively that the assumption is nearly correct, a reasonable first approximation, or entirely inaccurate.

different genera (for example, Daubin *et al.*, 2003) or even domains (for example, from bacteria to plant, Buchanan-Wollaston *et al.*, 1987; from yeast to bacteria, Heinemann and Sprague, 1989; or from archaea to bacteria, Nelson *et al.*, 1999).

2.4. SOME EXAMPLES OF NON-RANDOM PATTERNS OF RECOMBINATION

Here we describe a few examples where patterns of recombination are inhomogeneous within a genome and/or species. These examples have generally not come from systematic surveys of recombination and as such have an anecdotal quality to them but are valuable in defining the ways in which recombination rates do vary within natural populations.

2.4.1. Elements of the Genome Prone to Recombination

It has long been known that particular types of genes are prone to transfer. For example, antibiotic resistance genes are often found grouped in 'mobile gene cassettes', ready to be moved and transferred for example by integrons (Hall and Collis, 1995; Recchia and Hall, 1995). An example is provided by the *Vibrio cholerae* genome (Mazel *et al.*, 1998). In *Staphylococcus aureus*, genes responsible for the acquisition of methicillin resistance are all found in the same domain, which is horizontally transmitted and which seems to have been acquired from another bacterial species (Hiramatsu *et al.*, 2004). Genes for virulence factors are sometimes found in the same regions that contain antibiotic genes (Mazel *et al.*, 1998; Hiramatsu *et al.*, 2004). However, the lifestyle of pathogenic strains is usually completely different from the lifestyle of non-pathogenic strains so that acquisition of a large number of genes is required for the evolution of pathogenicity. This explains why virulence genes are usually found grouped together in virulence plasmids or in pathogenicity islands within the genome (Groisman and Ochman, 1997; Hacker *et al.*, 1997; Pickard *et al.*, 2003) which can then be transferred for example by transduction (Cheetham and Katz, 1995).

2.4.2. Sexual Isolation Due to Geographic or Ecological Subdivision

Some bacteria show non-random patterns of recombination due to physical proximity, which is comparable in effect to mating isolation found in eukaryotes. For example, strains of *H. pylori* recombine frequently with other strains found in the same human stomach during mixed infection. Because *H. pylori* is transmitted infrequently between people, there are large geographical differences in the frequencies of particular DNA sequence variants in different geographical locations, with strains progressively acquiring the genetic signature of their location by recombination with one another

(Falush *et al.*, 2003b). McCarthy *et al.* (2007) looked at strains of *Campylobacter jejuni* isolated from bovines and from chicken. Some lineages are found at high frequency in both reservoirs, indicating frequent exchange of bacteria between reservoirs. It is nevertheless possible to predict significantly more successfully than by chance alone which host the strains were isolated from by looking at alleles that they have imported. The observation of rapid import of locally prevalent alleles is biologically significant because it potentially allows the repeated evolution of host-adapted phenotypes in cosmopolitan species.

2.4.3. Formation of Virulent Lineages Associated With Elevated Recombination

There are several cases were virulent lineages have apparently undergone elevated rates of recombination. Most members of *Escherichia coli* are common and inoffensive inhabitants of mammalian intestines. There are, however, a few lineages that cause disease, resulting in about 1 million human deaths per year. Wirth *et al.* (2006) found that virulent *E. coli* had more mosaic ancestry, on average, than other strains and also had undergone a higher level of allelic divergence. The strain of *Vibrio vulnificus* responsible for an epidemic outbreak of disease in Israel in 1996 was shown to be a hybrid between two diverse populations of *V. vulnificus* that cause isolated cases of infection worldwide (Bisharat *et al.*, 2005). In *Staphylococcus aureus*, a pandemic methicillin-resistant lineage is also a hybrid due to a single very large genetic import (Robinson and Enright, 2004).

The invasive human-specific lineages of *Salmonella enterica*, Typhi and Paratyphi A, have converged with each other as a result of more than 100 distinct homologous recombination events, each several kilobases in size. These events happened in a rapid burst of recombination that appears to be entirely atypical of the normal level of recombination in *S. enterica* (Didelot *et al.*, 2007). There are two lines of evidence suggesting that this burst of recombination has already ceased. First, the segments of the genome that were exchanged consistently differ at about 2/1000 nucleotides, presumably reflecting the accumulation of mutations since import. Second, analysis of variation within Typhi shows a similar low level of recombination to that observed in other *S. enterica*. Specifically, In Typhimurium, 30 times more point mutations were detected than in imports from other strains of *S. enterica* (Didelot and Falush, 2007), whereas in Typhi, Roumagnac *et al.* (2006) found three imports (one potentially from Paratyphi A) and 83 single nucleotide changes that were probably due to point mutation. This

evidence for a burst of recombination is particularly interesting because it might be explained by an epidemic of bacteriophage infection affecting both Typhi and Paratyphi A. The data do not indicate whether exchange was approximately symmetric, Typhi imported DNA from Paratyphi A, or vice versa.

2.4.4. Virulent Clonal Lineages Within Recombining Populations

Despite the apparent role of recombination in the formation of virulent lineages, several such lineages show evidence of having become completely clonal in their mode of evolution, not importing any DNA. Examples include *Yersinia pestis*, the cause of the plague, which is a lineage of *Yersinia pseudotuberculosis* (Achtman *et al.*, 1999). This lineage is atypical in having two virulence plasmids that were presumably acquired recently, but variation within *Y. pestis* is entirely consistent with clonal descent (Achtman *et al.*, 2004). The closely related lineages of *Mycobacterium* causing tuberculosis in humans worldwide (*M. tuberculosis*), cattle (*M. bovis*), voles (*M. microti*), and humans living in Africa (*M. africanum*) are also apparently completely clonal (Kremer *et al.*, 1999; Smith *et al.*, 2003). However, they are closely related to the rare human pathogen *M. canetti*, which causes similar pathology in humans in East Africa that appears to undergo recombination (Gutierrez *et al.*, 2005).

2.4.5. Atypical Patterns of Observed Recombination Due to Selection

If natural selection promotes the spread of strains with particular genetic imports, observed rates of recombination may be atypically high even if the recombination rate itself is not elevated. Natural selection has obviously played a role in the frequent acquisition of antibiotic resistance and may also play a role in generating some of the other non-random patterns described earlier. An example where we have evidence for a role of selection rather than simply elevated recombination rates is in the acquisition of variants of the *tbpB* gene in subgroup A *Neisseria meningitidis* from *N. lactamica* and *Neisseria* spp. during epidemic spread. A much higher rate of genetic import is observed at the locus than at other loci, with most variants coding for a highly distinct version of the protein. These variants presumably allow immune escape in individuals that have been infected by strains carrying the original variant (Zhu *et al.*, 2001). The start and end points of the imported fragment differ between imports, with the fragment often stretching for

many kilobases either side of the *tpbB* allele (Linz *et al.*, 2000), showing that the elevated rate is not due to the propensity of a particular stretch of DNA for being frequently exchanged.

2.5. WHAT WE WANT TO KNOW

Notwithstanding the interesting examples just presented, our overall knowledge of variation in recombination rates compares unfavourably, for example, with our knowledge of variation in mutation rates within bacterial populations (Giraud *et al.*, 2001; Denamur and Matic, 2006) or variation in recombination rates between loci on the human genome (McVean *et al.*, 2004; Winckler *et al.*, 2005; Myers *et al.*, 2005).

For example we do not have answers to the following questions:

1. How much do neutral recombination rates vary across the genome?
2. What is the distribution of recombination import sizes?
3. How homology-dependent is recombination in vivo?
4. How common are bursts of elevated recombination, involving multiple imports within individual lineages?
5. How common are such bursts involving multiple imports between pairs of lineages?
6. How common are such bursts involving entire species?
7. Are mutators also recombiners?
8. What proportion of genetic imports are within species or between species?
9. To what extent is the import of non-homologous DNA associated with homologous recombination?
10. What are the ecological correlates of variation in recombination rate?
11. To what extent is the evolution of increased recombination during the acquisition of virulence a general phenomenon?
12. Why does recombination stop in some pathogen lineages?

We would also like to know about the importance of recombination for bacterial evolution:

1. What proportion of observed imports are under positive natural selection?
2. How much adaptation occurs through homologous recombination?
3. Do recombination rates correlate with the proportion of events that confer a selective benefit?

4. To what extent is the phenotypic and genetic cohesion of bacterial species maintained by recombination?
5. Does speciation occur because of the evolution of barriers to genetic exchange?

Most of these questions cannot be studied by looking at average properties across species, limiting the utility of several of the methods described earlier. Instead, information is required on the specific recombination events that have occurred in different lineages. The counting method of Feil *et al.* (2000) applied to MLST data provides a first approach to identifying these events but has several limitations. First, the method requires that genealogical relationships between strains be identified, which is difficult to do from MLST data (Didelot and Falush, 2007). The conservative approach that they use to identifying well-founded relationships (only analyzing single locus variants) means that no more than one event per strain is identified. Second, MLST fragments are too short to include entire imports, making it difficult to estimate fragment size, or indeed to robustly differentiate point mutations from imports. As a consequence the method provides little information overall, and almost none that is specific to particular lineages.

A more powerful approach is to estimate recombination events and genealogical relationships simultaneously. Recombination events cannot be identified unless we know about genealogical relationships, but at the same time inference of genealogical relationships in the presence of recombination is easier if the specific location of imports is known because the observed changes in the non-imported regions can be modelled assuming a molecular clock. Additionally, there is often uncertainty on both the genealogical tree and the location or presence of recombination events. Statistical approaches reach meaningful conclusions even in the presence of uncertainty about individual events. Our program, ClonalFrame, provides an example of a method able to jointly estimate clonal genealogies and imports (Didelot and Falush, 2007). ClonalFrame is a population genetics method in that it assumes a homogeneous recombination and mutation process across the population. However, inference is quite robust to this prior assumption, making it possible to detect changes in rates. It would also be possible to devise more complex parametric models to detect inhomogeneities.

Given appropriate analysis techniques, the amount of information provided by the data increases faster than linearly as the number of loci per strain increases because genealogical relationships become more accurate at the same as additional mutation and recombination events become detectable. Complete genome sequences provide both a large number of events and

information on the context and timing when each one occurred. For example, application of ClonalFrame to four genomes of *S. enterica* serovar Typhimurium identified 1733 mutation and 50 recombination events (Didelot and Falush, 2007). Of these, 141 and 4, respectively, occurred on an internal branch of the genealogical tree. The rate of mutation increased on one branch and decreased on another, showing that evolutionary processes are inhomogeneous even on this microevolutionary scale.

Several of our questions about variability in the recombination process could be addressed by using a similar approach applied to other serovars of *S. enterica* and contrasting the inferred patterns of imports. In order to investigate important rare events it may also be necessary to sample broadly across the species. For example if we could identify strains that shared common ancestry with either Typhi or Paratyphi A shortly before the burst of recombination that took place between these two lineages (Didelot *et al.*, 2007), we could apply the same methodology to place the recombination events in the context of clonal evolution. This could potentially provide reliable information on the timing of the events and also indicate the direction of genetic exchange. No close relatives of either serotype are known. If they do exist in the modern population, it is quite likely that they would only be identifiable using sequence datasets containing many more loci than traditional MLST. Thus, genomic analysis of specific lineages will not replace the need for good population surveys based on genomic datasets.

One area of analysis that has received little attention is the inference of the origin of specific recombination events. This is necessary to answer several of the questions above such as the relative importance of inter- and intraspecies recombination or the dependence of recombination on homology. ClonalFrame uses a highly simplistic model of recombination where imports introduce a constant proportion of nucleotide differences (Didelot and Falush, 2007). This approach has the advantage of not assuming that the population under study is closed but also means that a number of events are missed. In general, the problem of identifying the origin of events appears to be difficult because the genetic material may come from a source that is not in the sample, and also because recombination events between closely related strains are more difficult to detect, which can potentially introduce biases in the patterns of recombination observed.

Improved statistical methodology, together with large-scale sequence datasets becoming available, should provide a good understanding of the most complex and interesting patterns of bacterial recombination and microevolution and allow us to address general issues such as the importance of sexual process for genome adaptation and repair.

REFERENCES

Achtman, M. (2004). Population structure of pathogenic bacteria revisited. *Int J Med Microbiol*, **294**, 67–73.

Achtman, M., Zurth, K., Morelli, G., *et al.* (1999). *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proc Natl Acad Sci USA*, **96**, 14043–48.

Achtman, M., Morelli, G., Zhu, P., *et al.* (2004). Microevolution and history of the plague bacillus, *Yersinia pestis*. *Proc Natl Acad Sci USA*, **101**, 17837–42.

Anderson, E. S., and Williams. R. E. (1956). Bacteriophage typing of enteric pathogens and staphylococci and its use in epidemiology. *J Clin Pathol*, **9**, 94–127.

Baldo, L., J., Hotopp, C. D., Jolley, K. A., *et al.* (2006). A multilocus sequence typing system for the endosymbiont *Wolbachia*. *Appl Environ Microbiol*, **72**, 7098–110.

Bisharat, N., Cohen, D. I., Harding, R. M., *et al.* (2005). Hybrid *Vibrio vulnificus*. *Emerg Infect Dis*, **11**, 30–5.

Bjorkholm, B., Sjolund, M., Falk, P. G., *et al.* (2001). Mutation frequency and biological cost of antibiotic resistance in *Helicobacter pylori*. *Proc Natl Acad Sci USA*, **98**, 14607–12.

Brown, A., Feldman M., and Nevo, E. (1980). Multilocus structure of natural populations of *Hordeum spontaneum*. *Genetics*, **96**, 523–36.

Buchanan-Wollaston, V., Passiatore, J. E., and Cannon, F. (1987). The *mob* and *oriT* mobilization functions of a bacterial plasmid promote its transfer to plants. *Nature*, **328**, 172–5.

Cheetham, B. F., and Katz, M. E. (1995). A role for bacteriophages in the evolution and transfer of bacterial virulence determinants. *Mol Microbiol*, **18**, 201–8.

Chibani-Chennoufi, S., Bruttin, A., Dillmann, M. L., and Br, H. (2004). Phage-host interaction: an ecological perspective. *J Bacteriol*, **186**, 3677–86.

Clewell, D. B., Flannagan, S. E., and Jaworski, D. D. (1995). Unconstrained bacterial promiscuity: the Tn*916*-Tn*1545* family of conjugative transposons. *Trends Microbiol*, **3**, 229–36.

Daubin, V., Lerat, E., and Perriere, G. (2003). The source of laterally transferred genes in bacterial genomes. *Genome Biol*, **4**, R57.

Davison, J. (1999). Genetic exchange between bacteria in the environment. *Plasmid*, **42**, 73–91.

Denamur, E., and Matic, I. (2006). Evolution of mutation rates in bacteria. *Mol Microbiol*, **60**, 820–7.

Didelot, X., and Falush, D. (2007). Inference of bacterial microevolution using multilocus sequence data. *Genetics*, **175**, 1251–66.

Didelot, X., Achtman, M., Parkhill, J., Thomson N. R., and Falush, D. (2007). A bimodal pattern of relatedness between the *Salmonella* Paratyphi A and Typhi genomes: convergence or divergence by homologous recombination? *Genome Res*, **17**, 61–8.

Dingle, K. E., Colles, F. M., Falush D., and Maiden, M. C. (2005). Sequence typing and comparison of population biology of *Campylobacter coli* and *Campylobacter jejuni*. *J Clin Microbiol*, **43**, 340–7.

Dubnau, D. (1999). DNA uptake in bacteria. *Annu Rev Microbiol*, **53**, 217–44.

Falush, D., Kraft, C., Taylor, N. S., *et al.* (2001). Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. *Proc Natl Acad Sci USA*, **98**, 15056–61.

Falush, D., Stephens, M., and Pritchard, J. K. (2003a). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–87.

Falush, D., Wirth, T., Linz, B., *et al.* (2003b). Traces of human migrations in *Helicobacter pylori* populations. *Science*, **299**, 1582–5.

Falush, D., Torpdahl, M., Didelot, X., *et al.* (2006). Mismatch induced speciation in *Salmonella*: model and data. *Phil Trans R Soc B*, **361**, 2045–53.

Fearnhead, P., and Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics*, **159**, 1299–318.

Fearnhead, P., Smith, N. G., Barrigas, M., Fox, A., and French, N. (2005). Analysis of recombination in *Campylobacter jejuni* from MLST population data. *J Mol Evol*, **61**, 333–40.

Feil, E., Zhou, J., Maynard Smith, J., and Spratt, B. G. (1996). A comparison of the nucleotide sequences of the *adk* and *recA* genes of pathogenic and commensal *Neisseria* species: evidence for extensive interspecies recombination within *adk*. *J Mol Evol*, **43**, 631–40.

Feil, E., Maiden, M., Achtman, M., and Spratt, B. (1999). The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. *Mol Biol Evol*, **16**, 1496–502.

Feil, E. J., Smith, J. M., Enright, M. C., and Spratt, B. G. (2000). Estimating recombinational parameters in *Streptococcus pneumoniae* from Multilocus Sequence Typing data. *Genetics*, **154**, 1439–50.

Feil, E. J., Holmes, E. C., Enright, M. C., *et al.* (2001). Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc Natl Acad Sci USA*, **98**, 182–7.

Feil, E. J., Cooper, J. E., Grundmann, H., *et al.* (2003). How clonal is *Staphylococcus aureus*? *J Bacteriol*, **185**, 3307–16.

Feil, E. J., Li, B. C., Aanensen, D. M., Hanage, W. P., and Spratt, B. G. (2004). eBURST: Inferring patterns of evolutionary descent among clusters of related bacterial genotypes from Multilocus Sequence Typing data. *J Bacteriol*, **186**, 1518–30.

Fraser, C., Hanage, W. P., and Spratt, B. G. (2005). Neutral microepidemic evolution of bacterial pathogens. *Proc Natl Acad Sci USA*, **102**, 1968–73.

Giraud, A., Matic, I., Tenaillon, O., *et al.* (2001). Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut. *Science*, **291**, 2606–8.

Goodman, S. D., and Scocca, J. J. (1988). Identification and arrangement of the DNA sequence recognized in specific transformation of *Neisseria gonorrhoeae*. *Proc Natl Acad Sci USA*, **85**, 6982–6.

Goss, E. M., Kreitman, M., and Bergelson, J. (2005). Genetic diversity, recombination and cryptic clades in *Pseudomonas viridiava* infecting natural populations of *Arabidopsis thaliana*. *Genetics*, **169**, 21–35.

Griffiths, R., and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *J Comput Biol*, **3**, 479–502.

Groisman, E. A., and Ochman, H. (1997). How *Salmonella* became a pathogen. *Trends Microbiol*, **5**, 343–9.

Gutierrez, M. C., Brisse, S., Brosch, R., *et al.* (2005). Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*. *PLoS Pathog*, **1**, e5.

Guttman, D. S., and Dykhuizen, D. E. (1994). Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science*, **266**, 1380–3.

Hacker, J., Blum-Oehler, G., Muhldorfer, I., and Tschäpe, H. (1997). Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol Microbiol*, **23**, 1089–97.

Hall, R. M., and Collis, C. M. (1995). Mobile gene cassettes and integrons: capture and spread of genes by site-specific recombination. *Mol Microbiol*, **15**, 593–600.

Hanage, W. P., Fraser, C., and Spratt, B. G. (2005). Fuzzy species among recombinogenic bacteria. *BMC Biol*, **3**, 6–6.

Hanage, W. P., Spratt, B. G., Turner, K. M. E., and Fraser C. (2006). Modelling bacterial speciation. *Phil Trans R Soc B*, **361**, 2039–44.

Heinemann, J. A., and Sprague, G. F. (1989). Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. *Nature*, **340**, 205–9.

Hiramatsu, K., Watanabe, S., Takeuchi, F., Ito, T., and Baba, T. (2004). Genetic characterization of methicillin-resistant *Staphylococcus aureus*. *Vaccine*, **22** Suppl 1, 5–8.

Homan, W. L., Tribe, D., Poznanski, S., *et al.* (2002). Multilocus sequence typing scheme for *Enterococcus faecium*. *J Clin Microbiol*, **40**, 1963–71.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol*, **23**, 183–201.

Hudson, R. R. (2001). Two-locus sampling distributions and their application. *Genetics*, **159**, 1805–17.

Jolley, K. A., Kalmusova, J., Feil, E. J., *et al.* (2000). Carried meningococci in the Czech Republic: a diverse recombining population. *J Clin Microbiol*, **38**, 4492–8.

Jolley, K. A., Wilson, D. J., Kriz, P., McVean, G., and Maiden, M. C. J. (2005). The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in *Neisseria meningitidis*. *Mol Biol Evol*, **22**, 562–9.

King, S. J., Whatmore, A. M., and Dowson, C. G. (2005). NanA, a neuraminidase from Str*eptococcus pneumoniae*, shows high levels of sequence diversity, at least in part through recombination with *Streptococcus oralis*. *J Bacteriol*, **187**, 5376–86.

Kotetishvili, M., Kreger, A., Wauters, G., *et al.* (2005). Multilocus sequence typing for studying genetic relationships among *Yersinia* species. *J Clin Microbiol*, **43**, 2674–84.

Kremer, K., van Soolingen, D., Frothingham, R., *et al.* (1999). Comparison of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains: interlaboratory study of discriminatory power and reproducibility. *J Clin Microbiol*, **37**, 2607–18.

Kuhner, M. K., Yamato, J., and Felsenstein, J. (2000). Maximum likelihood estimation of recombination rates from population data. *Genetics*, **156**, 1393–401.

Le Minor, L. (1988). Typing of *Salmonella* species. *Eur J Clin Microbiol Infect Dis*, **7**, 214–8.

Linz, B., Schenker, M., Zhu, P., and Achtman, M. (2000). Frequent interspecific genetic exchange between commensal Neisseriae and *Neisseria meningitidis*. *Mol Microbiol*, **36**, 1049–58.

Lorenz, M. G., and Wackernagel, W. (1994). Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol Rev*, **58**, 563–602.

Maiden, M. C., Bygraves, J. A., Feil, E., *et al.* (1998). Multi-locus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci USA*, **95**, 3140–5.

Majewski, J., and Cohan, F. M. (1998). The effect of mismatch repair and heteroduplex formation on sexual isolation in *Bacillus*. *Genetics*, **148**, 13–8.

Majewski, J., and Cohan, F. M. (1999). DNA sequence similarity requirements for interspecific recombination in *Bacillus*. *Genetics*, **153**, 1525–33.

Majewski, J., Zawadzki, P., Pickerill, P., Cohan, F. M., and Dowson, C. G. (2000). Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation. *J Bacteriol*, **182**, 1016–23.

Maynard Smith, J., Smith, N., O'Rourke, M., and Spratt, B. (1993). How clonal are bacteria? *Proc Natl Acad Sci USA*, **90**, 4384–8.

Mazel, D., Dychinco, B., Webb, V. A., and Davies, J. (1998). A distinctive class of integron in the *Vibrio cholerae* genome. *Science*, **280**, 605–8.

McCarthy, N. D., Colles, F. M., Dingle, K. E., *et al.* (2007). Population genetic approaches to assigning the source of human pathogens: host associated genetic import in *Campylobacter jejuni*. *Emerg Infect Dis*, **13**, 267–72.

McGregor, K. F., Spratt, B. G., Kalia, A., *et al.* (2004). Multilocus Sequence Typing of *Streptococcus pyogenes* representing most known *emm* types and distinctions among subpopulation genetic structures. *J Bacteriol*, **186**, 4285–94.

McVean, G., Awadalla, P., and Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, **160**, 1231–41.

McVean, G. A., Myers, S. R., Hunt, S., *et al.* (2004). The fine-scale structure of recombination rate variation in the human genome. *Science*, **304**, 581–4.

Milkman, R., and Bridges, M. M. (1990). Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. *Genetics*, **126**, 505–17.

Millman, K. L., Tavaré, S., and Dean, D. (2001). Recombination in the *ompA* gene but not the *omcB* gene of *Chlamydia* contributes to serovar-specific differences in tissue tropism, immune surveillance, and persistence of the organism. *J Bacteriol*, **183**, 5997–6008.

Modrich, P., and Lahue, R. (1996). Mismatch repair in replication fidelity, genetic recombination, and cancer biology. *Annu Rev Biochem*, **65**, 101–33.

Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, **310**, 321–4.

Nelson, K. E., Clayton, R. A., Gill, S. R., *et al.* (1999). Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature*, **399**, 323–9.

Papke, R. T., Koenig, J. E., Rodriguez-Valera, F., and Doolittle, W. F. (2004). Frequent recombination in a saltern population of *Halorubrum*. *Science*, **306**, 1928–9.

Pickard, D., Wain, J., Baker, S., *et al.* (2003). Composition, acquisition, and distribution of the Vi exopolysaccharide-encoding *Salmonella enterica* pathogenicity island SPI-7. *J Bacteriol*, **185**, 5055–65.

Pritchard, J., Stephens, M., and Donnelly, P. J. (2000). Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–59.

Qiu, W.-G., Schutzer, S. E., Bruno, J. F., *et al.* (2004). Genetic exchange and plasmid transfers in *Borrelia burgdorferi* sensu stricto revealed by three-way genome comparisons and multilocus sequence typing. *Proc Natl Acad Sci USA*, **101**, 14150–5.

Recchia, G. D., and Hall, R. M. (1995). Gene cassettes: a new class of mobile element. *Microbiology*, **141**, 3015–27.

Robinson, D. A., and Enright M. C. (2004). Evolution of *Staphylococcus aureus* by large chromosomal replacements. *J Bacteriol*, **186**, 1060–4.

Robinson, D. A., Briles, D. E., Crain, M. J., and Hollingshead, S. K. (2002). Evolution and virulence of serogroup 6 pneumococci on a global scale. *J Bacteriol*, **184**, 6367–75.

Roumagnac, P., Weill, F.-X., Dolecek, C., *et al.* (2006). Evolutionary history of *Salmonella typhi*. *Science*, **314**, 1301–4.

Sarkar, S. F., and Guttman, D. S. (2004). Evolution of the Core Genome of *Pseudomonas syringae*, a highly clonal, endemic plant pathogen. *Appl Environ Microbiol*, **70**, 1999–2012.

Scally, M., Schuenzel, E. L., Stouthamer R., and Nunney, L. (2005). Multilocus Sequence Type system for the plant pathogen *Xylella fastidiosa* and relative contributions of recombination and point mutation to clonal diversity. *Appl Environ Microbiol*, **71**, 8491–9.

Schicklmaier, P., and Schmieger, H. (1995). Frequency of generalized transducing phages in natural isolates of the *Salmonella typhimurium* complex. *Appl Environ Microbiol*, **61**, 1637–40.

Schicklmaier, P., Moser, E., Wieland, T., Rabsch, W., and Schmieger, H. (1998). A comparative study on the frequency of prophages among natural isolates of *Salmonella* and *Escherichia coli* with emphasis on generalized transducers. *Antonie Van Leeuwenhoek*, **73**, 49–54.

Schouls, L. M., Reulen, S., Duim, B., *et al.* (2003). Comparative genotyping of *Campylobacter jejuni* by amplified fragment length polymorphism, multilocus sequence typing, and short repeat sequencing: strain diversity, host range, and recombination. *J Clin Microbiol*, **41**, 15–26.

Selander, R. K., and Levin, B. R. (1980). Genetic diversity and structure in *Escherichia coli* populations. *Science*, **210**, 545–7.

Smith, G. R. (1988). Homologous recombination in procaryotes. *Microbiol Rev*, **52**, 1–28.

Smith, N. H., Dale, J., Inwald, J., *et al.* (2003). The population structure of *Mycobacterium bovis* in Great Britain: clonal expansion. *Proc Natl Acad Sci USA*, **100**, 15271–5.

Spratt, B., Hanage, W., Li, B., Aanensen D., and Feil, E. (2004). Displaying the relatedness among isolates of bacterial species – the eBURST approach. *FEMS Microbiol Lett*, **241**, 129–34.

Suerbaum, S., Smith, J. M., Bapumia, K., *et al.* (1998). Free recombination within *Helicobacter pylori*. *Proc Natl Acad Sci USA*, **95**, 12619–24.

Tago, Y., Imai, M., Ihara, M., *et al.* (2005). *Escherichia coli* mutator Δ*polA* is defective in base mismatch correction: the nature of in vivo DNA replication errors. *J Mol Biol*, **351**, 299–308.

Tomb, J. F., White, O., Kerlavage, A. R., *et al.* (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature*, **388**, 539–47.

Vulic, M., Dionisio, F. Taddei, F., and Radman, M. (1997). Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc Natl Acad Sci USA*, **94**, 9763–7.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*, **7**, 256–76.

Weinbauer, M. G. (2004). Ecology of prokaryotic viruses. *FEMS Microbiol Rev*, **28**, 127–81.

Whitaker, R. J., Grogan, D. W., and Taylor, J. W. (2005) Recombination shapes the natural population structure of the hyperthermophilic archaeon *Sulfolobus islandicus*. *Mol Biol Evol*, **22**, 2354–61.

Winckler, W., Myers, S. R., Richter, D. J., *et al.* (2005). Comparison of fine-scale recombination rates in humans and chimpanzees. *Science*, **308**, 107–11.

Wirth, T., Falush, D., Lan, R., *et al.* (2006). Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol*, **60**, 1136–51.

Zahrt, T. C., and Maloy, S. (1997). Barriers to recombination between closely related bacteria: *MutS* and *RecBCD* inhibit recombination between *Salmonella typhimurium* and *Salmonella typhi*. *Proc Natl Acad Sci USA*, **94**, 9786–91.

Zhu, P., van der Ende, A., Falush, D., *et al.* (2001). Fit genotypes and escape variants of subgroup III *Neisseria meningitidis* during three pandemics of epidemic meningitis. *Proc Natl Acad Sci USA*, **98**, 15056–61.

Part II Mobile Genetic Elements in Bacterial Evolution

CHAPTER 3

Phage-bacterium Co-evolution and Its Implication for Bacterial Pathogenesis

Harald Brüssow

3.1. INTRODUCTION

Virulent phages such as *Escherichia coli* phage T4 are professional predators of bacteria. It is believed that this predator-prey relationship resulted in an evolutionary arms race in which bacteria developed anti-predation strategies against phages such as loss of receptor structures, restriction-modification systems, and abortive infection mechanisms. Temperate phages can lyse the bacterial host or alternatively integrate their DNA into the bacterial chromosome. As judged from the analysis of bacterial genomes, about a third of the bacteria might contain a prophage sequence. Temperate phages had thus a great impact on bacterial chromosome structure in general and the evolution of bacterial pathogenicity in special.

3.2. THEORETICAL FRAMEWORK FOR PHAGE-BACTERIUM INTERACTION

The peculiar life style of temperate phages makes them model systems to address a number of fundamental questions in evolutionary biology. The viral DNA undergoes different selective pressures when replicated during lytic infection cycles as compared to prophage DNA maintained in the bacterial genome during lysogeny. Darwinian considerations along with the selfish gene concept lead to interesting conjectures (Boyd and Brüssow, 2002; Brüssow and Hendrix, 2002; Brüssow *et al.*, 2004; Canchaya *et al.*, 2003, 2004; Lawrence *et al.*, 2001). One could anticipate that the prophage decreases the fitness of its lysogenic host by at least two processes: first by the metabolic burden to replicate extra DNA and second by the lysis of the

host after prophage induction. To compensate for these disadvantages one has to invoke the explanation that temperate phages encode functions that increase the fitness of the lysogen. According to the selective value of these postulated phage genes, the lysogenic cell will be maintained or even be over-represented in the bacterial population. An obvious selective advantage for the lysogenic host is the immunity (phage repressor) and super-infection exclusion genes of the prophage that protect the lysogen against extra phage infection. These genes are also of direct advantage for the prophage since they exclude super-infecting phage DNA from competing with the resident prophage DNA for the same host. Other phage genes might increase the fitness of the lysogenic host, frequently in rather unanticipated ways (lysogenic conversion genes). Classic examples include the non-essential phage λ genes *bor* and *lom* that confer serum-resistance and better survival in macrophages, respectively, to the *Escherichia coli* lysogen (Barondess and Beckwith, 1990). In these cases, the reproductive success of the lysogenic bacterium endowed with these new genes translates directly into an evolutionary success for the resident prophage. However, bacteria-phage interactions represent host-parasite relationships and therefore also an arms race leading to a highly dynamic genetic equilibrium. Gains from prophages carrying genes that increase host fitness are short-lived from a bacterial standpoint if the resident prophage ultimately destroys the bacterial lineage. In this way, prophages can be considered to be dangerous molecular time bombs that can kill the lysogenic cell upon their eventual induction (Lawrence *et al.*, 2001). One would therefore expect evolution to select lysogenic bacteria with mutations in the prophage DNA. Mutations that inactivate the prophage induction process avoid the loss of the lysogenic clone from the bacterial population. One would also expect that selection leads to large-scale deletion of prophage DNA in order to decrease the metabolic burden of extra DNA synthesis and a littering of the bacterial genomes with selfish DNA elements. One predicts furthermore that useful prophage genes (e.g., lysogenic conversion genes) are preferentially spared from this deletion process since their loss would actually decrease the fitness of the cell. It was also proposed that a high genomic deletion rate is instrumental in removing dangerous genetic parasites from the bacterial genome (Lawrence *et al.*, 2001). These deletion processes could explain why the bacterial genomes did not increase in size despite a constant bombardment with parasitic DNA over evolutionary time periods. The streamlined bacterial chromosome containing few pseudogenes might be the consequence of this deletion process of parasitic DNA.

3.3. REALITY TESTS

3.3.1. Growth Retardation of the Lysogen

A basic assumption of the foregoing hypothesis is that prophage DNA synthesis during each cell division is metabolically costly to the host. This seems an obvious bioenergetic statement. Yet, data published 30 years ago showed that λ lysogens of *E. coli* reproduce more rapidly than non-lysogens (Edlin *et al.*, 1975). However, a recent analysis came to a different conclusion (Chen *et al.*, 2005). The growth rates of a lambda lysogen and a non-lysogen did not differ on glucose, whereas on succinate a 30% increase in doubling times was observed in the lysogen. Such a growth disadvantage would quickly lead to an elimination of the lysogen from the bacterial population. However, the increase in the doubling time of the lysogen was not attributed to the additional cost of prophage DNA replication, but to a direct effect of the λ CI repressor on bacterial gene expression. The binding of CI to the promoter of the *pckA* gene resulted in a threefold transcriptional downregulation of this gluconeogenesis gene. The result is a lowering of the growth rate of the lysogen in energy-poor environments.

3.3.2. Inactivating Mutations in the Prophage

Bioinformatic analysis of bacterial genomes clearly showed a trend towards prophage inactivation, and induction experiments concur with this in silico interpretation. In a number of prophages from Gram-positive bacteria nonsense mutations in the portal protein gene or loss of the terminase gene affected the DNA packaging genes. In addition, several prophages showed DNA replication genes that were disrupted by a missense mutation, gene duplications, and rearrangements. In prophages from Gram-negative bacteria inactivating mutations seems to be enriched in the N antitermination genes, this process was most obvious for the prophages from *E. coli* O157 (Lawrence *et al.*, 2001). In accordance with this bioinformatic analysis, only one of the many λ-like prophages in *E. coli* O157 strain EDL933 and none in strain Sakai could be induced. However, such radical prophage inactivation is not a universal observation.

3.3.3. Reduction to Prophage Remnants

The frequent observation of prophage remnants in bacterial genomes suggests an ongoing prophage decay process. If the prophage decay process is

slow with respect to bacterial speciation, one would expect the conservation of closely related prophage remnants in different bacterial isolates from a given species. With a few exceptions (Canchaya *et al.*, 2002; Casjens, 2003) this was not observed. In fact, even closely related bacteria generally do not share prophage remnants: the phage remnants in different *E. coli* strains belonging to different serotypes are distinct. This observation suggests that the average time of acquisition and subsequent loss of prophages is shorter than the time needed for strain differentiation within a bacterial species. The nature of some phage genes suggests that phages are under deletion pressure. The *E. coli* phage P1 genome contains a toxin/antitoxin gene cassette. This genetic device is used by low-copy-number plasmids to prevent the loss from replicating bacteria. In fact, daughter cells not inheriting the plasmid encoding this toxin/antitoxin system are punished by the lethal effect of the long-lasting toxin protein, which is no longer neutralized by the short-lived antitoxin protein. One might suspect a peculiar role for this system in a phage persisting as a plasmid such as phage P1. However, a putative toxin/antitoxin cassette was also identified in a chromosomally integrated *Lactobacillus* prophage (Denou, personal communication).

3.4. EVOLUTION OF HUMAN BACTERIAL PATHOGENS REPRESENTS . . . 2-BILLION-YEAR-OLD TRENDS

Bacterial evolution has started long before the emergence of animals. In fact, bacteria (and Archaea) were the first and for some time the only inhabitants of earth. So early evolution only involved competition, genetic exchange, and selection between bacteria. One can assume that phages have also taken part in this early phase of evolution. If one judges from contemporary biology, bacteria became much more frequently the prey of phages than of predator bacteria (e.g., *Bdellovibrio*). At this time period of biological evolution, one might anticipate that bacteria and their phages were opposing biological forces engaged in a merciless arms race. Then the scene suddenly changed with the "birth" of the eukaryotic cell, which might perhaps go back to 2 billion years ago. Some biologists argue that not something fundamentally new was created with the first protists, but that they represented "only" new combinations of known pre-existing elements. A popular hypothesis speaks of the fusion of an archaeon with two bacteria. For bacteria these evolutionary new-comers were not only competitors for nutrients, but many protists probably discovered bacteria quite early as a source of food. In a more benign interaction the eukaryotic cell captured bacteria either as metabolic slaves or as endosymbionts, opening both partners in the latter case to new

evolutionary opportunities. For our current subject it is of greater interest that other protists became predators of bacteria. Now bacteria had to fight at two fronts: against phages and against protists. As evolution is never a one-way street, bacteria evolved powerful anti-predation measures.

One possible strategy of bacteria against grazing protozoa is surface masking. Amoebae recognize surface structures on the bacterial prey to set in the phagocytosis process. *Salmonella enterica* elaborates about 70 O types, chemically distinct forms of lipopolysaccharides (LPS), which decorate the surface of the bacterium. The classical interpretation is that LPS is a virulence factor important for escape from immune surveillance by the vertebrate host. However, *Salmonella* is not much exposed to the immune system. It is either in the gut or inside a cell. In contrast, experiments demonstrated selective feeding of gut amoeba on *Salmonella* belonging to different O serotypes. Possibly the primary ecological reason for the variability of the *Salmonella* O serotypes was to avoid protist grazing (Wildschutte *et al.*, 2004).

Once inside the food vacuoles of protists, bacteria had to opt for other escape strategies. They must change from pre-digestion to post-digestion anti-predation mechanisms – the most obvious being resistance to digestion. For example, a few minutes after ingestion of the cyanobacterium *Synechococcus*, flagellates reject the prey. They are apparently unpalatable, probably because of the proteinaceous S-layer with which the cyanobacterium surrounds itself (Matz and Kjelleberg, 2005). Under other conditions attack might be the best defense. Some bacteria elaborate toxins that have potent anti-protozoal activity (e.g., the pigment violacein produced by *Chromobacterium*). The predator flagellate *Ochromonas* does not distinguish between pigmented and non-pigmented bacteria. However, after digestion of as few as two pigmented bacteria, the predator lyses and releases its cell content. The ingested "altruistic" bacterium is dead (the toxin is not released from the intact cell), but the cytoplasm of the lysed protist feeds now the surviving bacterial population (Matz *et al.*, 2004).

Evolving along this latter line, some bacteria apparently learned in their turn to exploit protists as a rich food source. As protists evolved into multicellular organisms and later into increasingly complex animals, these predator bacteria had to co-evolve with their eukaryotic prey. Predator bacteria became thus bacterial pathogens of animals and plants. The predation pressure from pathogens was apparently so strong that animals mounted several lines of defense against them. One of the earliest and most effective defense tools against bacteria were phagocytic cells. They are used by primitive animals such as sponges sporting mobile amoebocytes as well as by humans. Now it is not a long step to hypothesize that tools initially deployed by bacteria

against protist predators were re-used by bacterial pathogens. This interpretation would nicely explain why such a surprising fraction of bacterial virulence factors addresses the problem of survival within the vacuoles of phagocytes.

Higher animals evolved an increasingly sophisticated immune system against invasion by pathogens. Along the arms race argument, bacterial pathogens had to follow this trend and were obliged to evolve proteins that paralyze or paradoxically over-activate the immune system.

What does all this mean with respect to the bacteria-phage relationship? This new antagonism between bacteria and protists (animals) also put the older antagonism between bacteria and phages into a new light. The interplay of three antagonizing biological systems created also the possibility of a two-partner coalition against the third protagonist, that is, cooperation between phages and bacteria against a protist prey. Pathogenic bacteria apparently "used" their viral predators to co-opt phage-carried genes to exploit a much larger gene pool in order to fight the immune system of birds and mammals. Whether eukaryotic viruses have an independent evolutionary origin or are derived from phages that learned to attack the eukaryotic host without the intermediate of bacteria cannot be decided at the moment. The structural similarities between the capsid structures of some phages and animal viruses could be an indication that such a phage-virus transition is not a necessarily far-fetched hypothesis.

3.5. . . . AND ADAPTATIONS OVER A SHORT TIME SCALE

Animals evolved over the past 800 to 600 million years, triggered into a radiation explosion during the Ediacaran/Cambrium by the rise in oxygen in the atmosphere, which provided a sort of "superfuel" for animal metabolism. Mammals proliferated massively only during the past 65 million years. Human-restricted pathogens such as *Streptococcus pyogenes*, *Shigella* spp., and the human-adapted *Salmonella* strains must have become adapted to their hosts within the time frame of human evolution of a few million years. This time frame calls for fast evolutionary processes. Genome alignments are here instructive. The comparison of *E. coli* with *Shigella* (at best a subspecies of *E. coli*, not a different genus as the name suggests) reveals DNA insertions (including many prophages), loss of genes, genome inversions, and translocations (some are flanked at one side by prophage sequences). Most prophages found in these two genomes differ in their DNA sequence, suggesting that they were acquired (and often degraded) after the separation of both bacterial lineages. However, at the DNA sequence level both

genomes could still be aligned over essentially the entire chromosomes. Or take two *E. coli* strains that belong to the same O157 serotype. The prominent role of prophages for strain differentiation can be directly gleaned from the dotplot alignment. In fact, most of the current evidence for the involvement of phages in shaping bacterial genomes, bacterial fitness, and host-pathogen interactions deals with events at the level of intra-species differences. The conspicuous role of horizontally transferred DNA for the evolution of pathogens is probably the consequence of the time available for bacteria to adapt to humans as a new niche. The balancing of the pathogen-host interaction has been relatively short and will therefore be enriched for quick adaptation processes.

3.6. EVOLUTIONARY OPTIONS

Bacterial evolution requires modifying old functions and developing new ones. Nucleotide exchange and nucleotide insertions or deletions are the most frequent events. Mutation rates in bacteria are generally in the range of 10^{-6} to 10^{-9} per nucleotide per generation. In addition, exchanges, deletions, and insertions of larger stretches of DNA are occurring at appreciable frequency. These mechanisms are common to all living organisms and allow modification of existing functions in order to optimize fitness in an existing niche or the adaptation to a new niche. Virulence and fitness factors can thus evolve "vertically" by gene duplication, mutation, and even by gene disruption. However, this needs quite long time periods to achieve results, especially when a complex combination of genes is required for a new niche. In contrast to many higher eukaryotes, bacteria have no sexual life cycles to facilitate exchange of alleles within a population. Vertical evolution is thus too slow to drive the evolution of bacterial pathogens over a million years, not to speak of decades. However, lateral gene transfer can fill in this ticket. Combination of new gene constellations by the permutation principle is readily achieved by mobile DNA elements such as phages. In bacteria, in contrast to eukaryotes, entire functional units can be imported from other sources and are principally not restricted by species barriers. The transferred DNA can range in size from less than 1 to more than 100 kb. It can encode entire metabolic pathways or complex surface structures. These genes can be taken up as naked DNA or transferred in the form of plasmids, conjugative transposons, or phages.

It thus seems that bacteria evolve with two gears. The slow mode is based on the usual mechanisms of vertical evolution mediating a step-by-step genetic adaptation to their approximate environment. Lateral gene

transfer can be regarded as the fast mode of evolution (years to decades time range). Most of these gains might be ephemeral and are as easily lost (Lawrence, 1997; Lawrence and Ochman, 1997). What counts is a momentary selective advantage over competing bacteria especially in environments that are quickly changing. This changing environment can be the body surfaces of new host species that are rapidly proliferating in the ecosphere (e.g., mankind) or settings with unusual host densities (animal farming, human urbanization). In fact, overpopulation, international travel of persons and transportation of food and livestock, social upheaval, and wars create enormous possibilities for microbes that can exploit them. The impact of lateral gene transfer is too obvious for short-term bacterial evolution, such as the spread of antibiotic resistance genes. Phage DNA fulfils a number of criteria for being an ideal vehicle for lateral gene transfer. Different combinations of mobile DNA can be explored; suitable combinations are maintained and further developed, leading to genotypes that suddenly fill old and newly created niches. New diseases such as "flesh eating" *S. pyogenes* or the "hamburger disease" caused by the emerging food pathogen *E. coli* O157:H7 apparently evolved over decades. Phages visibly played an important role in these short-term adaptation processes. The effect of this quick process on the long-term evolution of bacteria is less certain. It seems that only very small amounts of prophage DNA are fixed into the bacterial chromosome.

3.7. PROPHAGES CONTRIBUTE TO THE GENETIC INDIVIDUALITY OF A BACTERIAL STRAIN

Data from comparative bacterial genomics highlight the importance of lateral DNA transfer in microbial evolution (Bushman, 2002). Based on a variety of criteria such as sequence matches with other organisms, G+C-content, codon usage, association with mobile DNA elements, and proximity to tRNA genes, it has been estimated that some bacteria capture and fix DNA at a rate of at least 16 kilobases per million years (Lawrence and Ochman, 1998). An interesting case is provided by *E. coli*: Genomic comparisons between the pathogenic *E. coli* O157 strain EDL933 and the laboratory *E. coli* strain K12 revealed 4.1 Mb of common chromosome backbone sequence and 1.3 Mb of O157-specific and 0.5 Mb of K12-specific DNA (Ohnishi *et al.*, 2001). Approximately half of the O157-specific DNA was clear-cut mobile DNA, mostly prophage DNA (Perna *et al.*, 2001). These observations led to the distinction of a conserved core genome sequence, which is shaped by the mechanisms of vertical evolution and a variable part of the genome, which is dominated by processes of horizontal evolution. Interestingly, prophage comparisons

between strains sharing a very recent common ancestor already showed modular exchange reactions (Makino *et al.*, 1999) suggesting that prophages are a highly dynamic part of bacterial genomes.

In a *Salmonella enterica* serovar Typhi and Typhimurium comparison, two chromosomal inversions and 12 larger alignment gaps were identified. Nine of the gaps can be traced to prophages or prophage remnants. The gaps represent either prophage insertion/deletion events or prophage replacements at a given chromosomal position. Similar observations can be cited for Gram-positive bacteria. The alignment of two *Streptococcus agalactiae* serotypes showed about a dozen small gaps. The gaps corresponded nearly exclusively to mobile DNA. In some *Staphylococcus aureus* strain comparisons, prophages were the major contributors to variability (Kuroda *et al.*, 2001), while in other comparisons they competed with genomic islands and transposons for this role (Baba *et al.*, 2002). An extreme case is presented by *Streptococcus pyogenes* in which practically all major gaps in the alignment of different M serotypes could be traced to prophage integration events (Smoot *et al.*, 2002). Since all *S. pyogenes* and many *S. aureus* prophages encode proven or suspected virulence factors, the prophage-induced diversity between the strains might explain why such chromosomally similar bacteria can cause such a wide range of clinical symptoms (Beres *et al.*, 2002).

It is not rare that different prophages from the same lysogen share DNA sequence identity. As these regions are targets for homologous recombination, it was predicted that prophages mediate rearrangements of bacterial chromosomes (Brüssow and Hendrix, 2002). Several genome alignments support this prediction. For example, a Japanese *S. pyogenes* M3 strain differed from an American M3 isolate mainly by two sequential DNA inversions (Nakagawa *et al.*, 2003). One inversion occurred across two prophages. Genomics analysis suggested that the crossover point was located in the lysis modules. Since the lysogenic conversion genes from *S. pyogenes* prophages are encoded downstream of the lysis genes, the cross-over results in a reshuffling of virulence genes between prophages. This might allow a wider horizontal spread of these genes, because they become associated with phages having potential new host ranges and belonging to new immunity groups. This flexibility extends the possibilities of conversion gene permutation in polylysogenic hosts such as *S. pyogenes*. A spectacular case of a likely genome rearrangement is presented by the Gram-negative plant pathogen *Xylella fastidiosa*. Two pathovars shared DNA sequence identity essentially over the entire genome but showed a very complex alignment pattern suggestive of three successive chromosomal inversion events across prophage DNA (van Sluys *et al.*, 2003).

3.8. BECOMING A PATHOGEN

3.8.1. From Virulence to Fitness Factors

How do bacteria adapt to the lifestyle of a pathogen? At first glance, mammalian hosts are ecological niches like any other. However, in some respects they are difficult niches. In addition to the normal challenges encountered in non-living ecological niches, mammals have defenses, which have been shaped by co-evolution with microbes. This includes simple physical barriers such as the dead cell layers of the skin or mucus-covered epithelia, or the elaboration of antimicrobial peptides, iron sequestering mechanisms, and immune responses. The factors and mechanisms that pathogens have evolved to circumvent these defenses are termed virulence factors. Virulence factors come in a great variety and include factors that neutralize defenses of the host and factors that help to engage, subvert, or destruct cells of the animal host.

Generally, the interaction of a pathogen with the host is a multistage process, which includes searching an entry site, targeting a place for multiplication in the body of the host, taking precautions for persistence in the original host, or finding ways to the next host. Survival or multiplication in the environment and transmission to the next host, which might include invertebrate vectors, certainly affect the overall success of a pathogen. Virulence factors might thus become a special case of more generalized "fitness factors."

3.8.2. Phage-encoded Virulence Factors

The acquisition of prophages is an irrelevant process for the evolution of pathogenic bacteria if phages do not transfer useful genes to the lysogen. From evolutionary reasoning, we should therefore not be surprised that phages carry important virulence factors for bacterial pathogens. That phages carry key virulence factors was discovered long ago. Two examples are the phages γ and C1, which encode key virulence factors, namely toxins, of *Corynebacterium diphtheriae* and *Clostridium botulinum*, respectively (Barksdale and Arden, 1974; Freeman, 1951). The discovery of the cholera toxin-encoding phage CTXΦ (Waldor and Mekalanos, 1996) was one of the landmarks of this new phage research. It was becoming increasingly clear that phages play an important role in the evolution and virulence of many pathogens. Why should phages encode toxins that act on eukaryotic cells, which they cannot infect under natural conditions? As far as we know these

genes do not play a role in the life cycle of the phages. A key to this question is provided by a list of phage-encoded fitness factors, which is rapidly growing and now comprises a wide range of different genes. This includes many toxins, superantigens, LPS modifying enzymes, type III effector proteins, detoxifying enzymes, hydrolytic enzymes, and proteins conferring serum resistance or resistance against phagocytosis. In a few cases, phage tail genes seem to have developed dual functions and serve also as adhesion proteins for bacterial host attachment (Bensing *et al.*, 2001). If one reads through this list, one cannot escape the impression that these are just those proteins needed by bacteria that want to make a living from animal hosts. It seems that bacteria have used their ancient enemies for their purpose by exploiting their gene transfer capacities.

3.9. PHAGES CARRYING VIRULENCE GENES

3.9.1. Lambdoid Siphoviridae

Virulence genes are carried by many different types of phages, but lambdoid Siphoviridae are clearly a preferred carrier. In lambdoid prophages from low G+C content Gram-positive bacteria the preferred location of lysogenic conversion genes (LCG) is downstream of the phage lysin gene in the vicinity of *attR*, the right attachment site. Two reasons explain this position. These extra genes belong frequently to the few genes transcribed from the prophage genome. Because the transcription of late genes and especially the lysis genes would upset the quiescent prophage state, it is safe to place the virulence genes at the transition zone from prophage DNA into the bacterial DNA. Transcription would then either run into bacterial genes or in case of an opposite orientation into the late phage genes as an anti-messenger. The second reason for this location of the extra genes might be linked to the acquisition process of LCG. Following an imprecise excision process, adjacent bacterial DNA accidentally remains joined to the excised prophage DNA and is packaged into the phage particle. Sometimes these extra genes differ in G+C content from the surrounding DNA, suggesting gene transfer from a rare host differing in G+C content (Ferretti *et al.*, 2001). Also, in high-G+C-content Gram-positive bacteria such as *Corynebacterium diphtheriae* the diphtheria toxin is also encoded on a strikingly similar lambdoid prophage downstream of the putative tail fiber genes and next to *attR*. The expression of the diphtheria toxin gene is regulated via the bacterial DtxR transcriptional regulator that binds in an iron-dependent way to operators of many *C. diphtheriae* genes. The second location of virulence genes in prophages

from low-G+C-content Gram-positive bacteria is in the lysogeny module, specifically between the phage repressor and the integrase genes next to the *attL* site. Since this region is constitutively transcribed in the prophage state, it is a logical place for these extra genes.

In lambdoid prophages from Gram-negative bacteria, virulence genes are also frequently found near the lysis genes. However, in these phages the lysis genes are centrally placed ahead of the DNA packaging genes. Other preferential insertion sites are downstream of the Q anti-terminator and the N anti-terminator genes (Boyd and Brüssow, 2002). However, many extra genes were also detected at variable prophage genome locations, sometimes in the middle of phage modules. The term "morons" (more DNA) was coined for these extra DNA elements. They represent transcription units with their own promoters and terminators that are regulated independently from the rest of the prophage (Wagner *et al.*, 2001b). Some of the LCG were shown to respond to environmental cues (Broudy *et al.*, 2001; Wagner *et al.*, 2001a). In fact, when bacteria were grown under conditions that mimicked pathological conditions (Smoot *et al.*, 2001) or when they were grown in infected animals (Dozois *et al.*, 2003), lambdoid prophage genes belonged to the most prominent genes changing the expression level.

3.9.2. Inoviridae

The cholera toxin is encoded on the genome of a filamentous phage, which became the focus of much research (McLeod *et al.*, 2005; Waldor and Friedman, 2005). This prophage is in many respects atypical. Its *E. coli* cousins M13 and fd do not integrate into the bacterial chromosome and are maintained as plasmids. CTXΦ from *Vibrio cholerae* achieves integration via the host-encoded recombinases XerC and XerD (Huber and Waldor, 2002). CTXΦ phage does not lyse the host cell – the phage is secreted through a chromosomally encoded outer membrane pore. Lysogeny has thus only minimal effects on the cellular growth rate. The cholera toxin (CT), a classical AB toxin, leaves the cell through the same EspD type II secretion system as the phage. The B subunit of CT binds to enterocytes and transports the catalytic A subunit into the host cell cytoplasm. The subunit A triggers a signaling cascade which leads to chloride and water efflux into the intestinal lumen. Purified CT, when given orally to volunteers, produced watery diarrhea, the hallmark of cholera. However, the expression level of CT during disease is still a matter of debate. The two CT genes *ctxA* and *ctxB* are found at the 3′ end of the short genome of this filamentous phage. Most of the CT transcripts are initiated at a promoter just upstream of the *ctxA,B* genes

and are under the control of cellular transcription factors (ToxR, ToxT, and TcpPH). Transcription can also be initiated from a phage promoter ahead of the *rstA* gene next to the *rstR* phage repressor at the 5′ end of the CTXΦ genome. Toxigenic *V. cholerae* strains usually contain tandem CTX prophages with an RS1 element interspersed between the two CTX prophages. RS1 looks like a prophage genome remnant that depends on CTXΦ for coat and assembly proteins. The prophage locus encoding the CT is thus likely the result of multiple integration events and might thus be the reason why filamentous phages – despite the prominence of the CT paradigm – are not a frequently used carrier for bacterial virulence factors.

3.9.3. Other Phages Less Commonly Used as Virulence Gene Carriers

Despite its prominence for the history of molecular biology, Mu-like phages that replicate via transposition are rare in *E. coli*. Genome sequencing projects have now shown Mu-like prophages in a number of other bacteria and recently a Mu-like Myovirus prophage from *Burkholderia* was described carrying potential virulence factors (Summer *et al.*, 2004). The *B. cepacia* complex includes plant and animal pathogens and plays a role in cystic fibrosis patients. Gene 8, located in the lysogeny and transposition module of such a *Burkholderia* Mu-like prophage, encodes a protein that resembles an *Aeromonas* protein involved in the secretion of a toxin. Gene 53, located at the very end of the same Mu-like genome, encodes a 3-O-acyltransferase, which might be involved in acetylation of the LPS O-antigen. Likewise, the P2-like Myovirus ΦCTX from *Pseudomonas aeruginosa* encodes a pore-forming cytotoxin CTX at the very end of its genome. It appears as if the *ctx* gene has jumped just in between *cosL* at the genome end and the first ORF, encoding the capsid protein Q (Nakayama *et al.*, 1999). Another important prophage-encoded toxin is the botulinum type C neurotoxin. This was historically one of the first toxins associated with prophages. Botulinum neurotoxin (BoNTX) produced by the food pathogen *Clostridium botulinum* is one of the most poisonous substances known. It is secreted as a progenitor toxin associated with non-toxic hemagglutinins, which protect the toxin from destruction by gastric juice. The toxin is adsorbed through the upper intestinal tract and binds to the presynaptic membrane, where it interferes with the release of the neurotransmitter. The genes encoding the entire progenitor toxin protein complex are inserted as a contiguous gene cluster into the DNA transaction module of one of the largest temperate phages, sporting a 185 kb genome (Sakaguchi *et al.*, 2005). It is a Siphovirus distantly related to

the *Bacillus subtilis* prophage SPβ. It persists as a low copy circular plasmid prophage, which leads to an instability of the lysogenic state that was called by virologists "pseudo-lysogeny." Non-toxigenic derivatives emerge thus at high frequency. The insertion of the toxin gene cluster into a central phage module is not a particularity for this toxin DNA element since the prophage has in addition suffered the acquisition of 12 Insertion Sequences. Furthermore, not all *C. botulinum* strains carry related neurotoxins on prophages, in some they are encoded on the bacterial chromosome.

Phages do not only confer genetic mobility to bacterial genomes, they also lead to recombinations between the genomes from different phage classes. Chimeric phages between Siphoviridae and Myoviridae are frequent observations. In *Shigella* and *Salmonella* such hybrid phages also carry virulence genes (e.g., O serotype-converting enzymes) near *attL*, the left attachment site (Allison *et al.*, 2002).

3.10. PHAGES MEDIATE GENE TRANSFER BETWEEN BACTERIAL GENOMES

3.10.1. Bacteria Exploiting Phages for Their Ends: Gene Transfer Agents

Tailed phages are the most efficient gene transfer particles developed in evolution. They represent densely compacted DNA (Zhang *et al.*, 2000) encased in a protective protein shell, the phage head (Conway *et al.*, 2001). To this remarkable DNA storage device is added an equally efficient DNA transfer device, the phage tail. This structure assures both the specific recognition of the appropriate host cell and the guided injection of the phage DNA into the bacterial cell (Kanamuru *et al.*, 2002; Molineux, 2001). Some bacteria have apparently learned to use phages for their own purposes. One obvious use is as gene transfer particles containing bacterial DNA instead of phage DNA in the capsid. In this way bacterial DNA can be efficiently shuttled within a bacterial species. We do know a number of such "tamed" phages. The defective *B. subtilis* prophage PBSX has maintained the capacity to build a size-reduced phage head, into which 13-kb fragments of random bacterial DNA are packaged. A prophage remnant of *Rhodobacter capsulatus* acts as a gene transfer agent, transporting random 4.5 kb-sized fragments of bacterial DNA between cells in the stationary phase (Lang and Beatty, 2000). The bacteria control the DNA exchange by a two-component signal transduction system. Comparable particles are widely distributed among *Rhodobacter* and other α-proteobacteria. A similar particle transferring random 7.5-kb fragments

of bacterial DNA was identified in the spirochete *Brachyspira hyodysenteriae*, a swine pathogen. The protein components of this particle are encoded on a 16-kb genome segment, located on the bacterial chromosome that comprises head and tail genes plus a lysis cassette of a standard Siphovirus. Early phage genes and DNA packaging genes were not detected. This prophage has thus lost the capacity of self-replication and the virions are non-infectious. The population structure of its host is shaped by a high level of genetic recombination suggesting that the ex-prophage is actually used as gene transfer agent by its host (Matson *et al.*, 2005).

A particularly interesting case is the 15-kb pathogenicity island SaPI1 from *S. aureus* encoding the Tst toxin involved in toxic shock. In cells infected with *S. aureus* phage 80α, SaPI1 is excised from the chromosome, replicates autonomously, and interferes with phage growth by directing the encapsidation of its own DNA into specially tailored small phage 80α heads. Upon phage-mediated transfer to a recipient organism, SaPI1 integrates into the new host chromosome by means of its own integrase (Ruzin *et al.*, 2001).

Bacteria have also exploited other properties from phages. In *Pseudomonas aeruginosa* two phage tail gene clusters have developed into bacteriocins (Nakayama *et al.*, 2000). Bacteriocins that resemble phage tail structures are known from a number of other bacteria (Strauch *et al.*, 2001).

3.10.2. Transduction

Resolvase-type integrases from phages of Gram-positive bacteria have no requirements for cofactors, facilitating the integration of prophages into heterologous hosts (Groth *et al.*, 2000). If a prophage is imprecisely excised from the heterologous host, small segments of flanking bacterial DNA can be co-packaged with the phage DNA and transferred to the original host (specialized transduction). Some phages such as *Salmonella* phage P22 or coliphage Mu occasionally commit the error of packaging even a headful of bacterial DNA instead of phage DNA. Upon infection of the next host, this bacterial DNA can be incorporated into the bacterial chromosome (generalized transduction). Marine microbiologists draw our attention to the enormous extent of phage-mediated DNA transfer in the ocean (Williamson *et al.*, 2002). They noted that viruses outnumber bacteria in the open ocean by a factor of 10 (Wommack and Colwell, 2000). In view of the large volume of the world's oceans and the high titer of phage particles of 10^7 per ml of seawater, or probably 10^{30} particles, phages are the most abundant biological entities on earth. If one anticipates a transduction frequency of 10^{-8} per plaque-forming unit (Jiang and Paul, 1998), it was calculated that phage-mediated gene transfer

takes place at the incredible rate of about 20 million billion times per second in the oceans (Bushman, 2002).

3.11. THE WIDE REACH OF PHAGE GENES

3.11.1. Induction of Insect Pathologies – The Case of the Cytolethal Distending Toxin in Arthropods: *Hamiltonella*

Unconventional systems provide perhaps the best demonstration for the advantage of being infected. As an example I will here quote a biological Russian doll consisting of a prey insect (a plant sap-sucking aphid), a predator insect (a parasitoid wasp laying eggs into the aphid), and two bacterial endosymbionts of the aphid. The primary endosymbiont is *Buchnera*, whose metabolism helps the aphid to cope with dietary deficiencies imposed by its exclusive feeding on plant food streams not adapted for the nutritional needs of an animal. The secondary endosymbiont is *Hamiltonella defensa*. Experiments with aphids demonstrated that carriage of *Hamiltonella* did not prevent attack of aphids by the parasitoid. However, the eggs of the parasitoid were less likely to develop in the infected aphids than in aphids lacking *Hamiltonella* (Oliver *et al.*, 2003). Subsequent experiments assured that the resistance phenotype was associated with the genotype of the secondary endosymbiont and not that of the aphid. However, the level of resistance varied from 19 to 100% (Oliver *et al.*, 2005). Sequence analysis of *Hamiltonella* identified the last partner in this Russian doll: the secondary endosymbiont was infected with a bacteriophage, which resembled a P22-like lambdoid phage known from another aphid symbiont carrying a Shiga-toxin gene. In the *Hamiltonella* prophage the Shiga-toxin encoding *stx* cassette was replaced by a cassette encoding the cytolethal distending toxin (CDT) (Moran *et al.*, 2005). This protein belongs to a newly discovered toxin family, which blocks mammalian cells in the G2 phase of the cell cycle. CdtB is the active A subunit of the tripartite CDT holotoxin. The heterodimeric B subunit is required for the delivery of CdtB into the eukaryotic target cell. The protein shows features of a type I deoxyribonuclease. Indeed, transient expression of this protein in cultured cells causes limited DNA damage that results in chromatin disruption, cytoplasmic distention, and ultimately cell cycle arrest (Lara-Tejero and Galan, 2000). CdtB was quoted as an example how bacterial pathogens have co-evolved with their eukaryotic hosts to manipulate host cell functions precisely in order to trigger only limited damage for an optimal exploitation of their hosts (Lara-Tejero and Galan, 2002). CdtB

is found in various pathogens such as *Campylobacter*, *Shigella*, *Salmonella*, *Haemophilus*, and *Helicobacter* (Bielaszewska *et al.*, 2005) and notably also in Shiga toxin-producing *E. coli* STEC (Bielaszewska *et al.*, 2004).

We have to consider long chains of interactions involving phage, bacteria and insects if we want to disentangle webs of co-evolution (Moran *et al.*, 2005). Phage genes thus have a wide reach. The flexibility of phage genomes to incorporate different toxin genes adds substantially to their value as gene transfer particles that provide possibilities for evolution in the fast lane.

3.11.2. Another Tripartite Phage-bacterium-arthropod Relationship: *Wolbachia*

The bacterium *Wolbachia* lives in the reproductive cells of its arthropod host where it causes a particular phenotype that has fascinated biologists: cytoplasmic incompatibility. *Wolbachia*-infected females yield progeny with both infected and uninfected males, resulting in infected animals. In contrast, when uninfected females mate with infected males, the embryo dies. This incompatibility favors the spread of *Wolbachia* through the arthropod population to high prevalence level. This reproductive bacterial parasitism made *Wolbachia* a very successful organism, which is now a maternally transmitted bacterium in 20% of all arthropods. Since insects comprise 85% of all animal species, it can be argued that *Wolbachia* is one of the most successful intracellular parasites on earth. The *Culex pipiens* mosquito group is a *Wolbachia*-infested species. *C. pipiens* showed a complex pattern of crossing sterilities, but no *Wolbachia* strain variation. Sequencing of *Wolbachia* pointed, however, to a variable region within its prophages. This prophage region encodes ankyrin proteins known as transcription factors that modify cell-cycle proteins and that mediate protein interactions with the cytoskeleton. Ankyrin genes were also detected in phage-like particles found in *Culex* reproductive cells. Finally, a *Wolbachia* strain that no longer expressed incompatibility showed a stop codon in the prophage-encoded ankyrin gene (Sinkins *et al.*, 2005). However, the jury is still out whether this prophage gene is the motor for the sweep of *Wolbachia* through insect populations. Other researchers deduced a different scenario: Here phage infects and kills *Wolbachia*, thereby reducing *Wolbachia*'s population density. Instead of being the motor of cytoplasmic incompatibility, here phage is a break in this tripartite relationship, keeping at bay the toxic effect of bacterial endoparasitism for the insect (Bordenstein *et al.*, 2006).

3.11.3. Anti-feeding Genes: *Serratia*

Serratia entomophila is the causal agent of amber disease of a beetle that is an important pasture pest in New Zealand. This bacterium was developed into a commercial biopesticide. The larvae of the beetle show cessation of feeding, clearance of the gut, amber discoloration, and death as the disease progresses. The anti-feeding genes were genetically attributed to a defective prophage consisting of tail sheath and baseplate genes (Hurst *et al.*, 2004). Two features make the *Serratia* prophage remnant unusual. It resides on a plasmid and it is preceded by a holin-lysin cassette. The release of the *Serratia* anti-feeding protein complex might thus be under the control of the prophage remnant.

3.11.4. *E. coli* and *Salmonella* Phages: Takeover of the Enterocyte

Enteropathogenic *E. coli* (EPEC) and enterohemorrhagic *E. coli* (EHEC) are closely related pathogens that colonize the intestinal mucosa. Both bacteria cause specific lesions on the enterocyte, which comprise attaching and effacing lesions. These lesions are characterized by intimate bacterial adhesion, reorganization of the host cytoskeleton, cell cycle arrest, opening of the tight junctions between the cells, and destruction of the brush border microvilli. Both pathogens encode a type III secretion system (T3SS) and effector proteins on a chromosomally encoded pathogenicity island called "locus of enterocyte effacement" (LEE), which is not found in commensal gut *E. coli* or pathogenic *E. coli* strains causing other types of disease. T3SS are molecular needles that inject bacterial virulence factors directly into the host cell. For example, EPEC and EHEC inject the bacterial Tir protein into the enterocyte, which inserts into the plasma membrane of the gut cell and acts there as a receptor for intimin, a bacterial adhesin. EHEC strains like O157:H7 cause the feared hemolytic uremia syndrome. These are food pathogens that emerged worldwide over the past 20 years. Molecular epidemiologists demonstrated how this pathogen evolved in a series of steps from a non-toxigenic O55:H7 clone, a major EPEC pathogen of childhood diarrhea. The ancestral clone first acquired the entire O-antigen gene cluster by horizontal transfer, then the two Shiga toxin-encoding phages were acquired, followed by many gains and losses of prophages (Wick *et al.*, 2005). The sequenced O157:H7 strains carry an astonishing load of prophages. If you look at a genome map of these strains, which depicts 18 prophage sequences in the epidemic Sakai strain, you get the impression of a bacteriophage in bacterial disguise.

The implication of the prophage sequences for the pathogenesis of EHEC infections has been intensively explored over the last years. Two of these prophages encode Shiga-like toxins. As the name suggests, these proteins resemble the toxins elaborated by *Shigella*, the cause of dysentery. Its mode of action was biochemically described. Shiga toxins that occur as two groups with distinct antigenic properties, Stx1 and Stx2, are AB toxins. They consist of a B pentamer that binds to glycolipids of the target cell and an A protein that inhibits protein synthesis in susceptible cells. Its biochemical mode of action was characterized: The A protein is an N-glycosidase that removes a single base from the 28 S rRNA of the eukaryotic ribosome (Saxena *et al.*, 1989). The orally applied cytotoxin is lethal to animals at low doses. It was not apparent how this protein could benefit the survival of the bacterium. This function was now revealed: Shiga toxins enhance the capacity of EHEC O157:H7 to adhere to epithelial cells and to colonize the intestine of mice (Robinson *et al.*, 2006). The mode of action is probably indirect: Shiga toxin enhances the surface expression of nucleolin (the primary function of this nuclear protein is in ribosome biogenesis). Nucleolin interacts with intimin, the bacterial adhesin, which normally interacts with the bacterial Tir receptor in the eukaryotic plasma membrane. Interestingly, EHEC Tir in contrast to EPEC Tir is not sufficient to initiate F-actin polymerization, resulting in stress fiber formation. EHEC Tir is not phosphorylated at a critical tyrosine residue and it does not stimulate the Nck-mediated signaling cascade. The activation of N-WASP by Nck therefore requires an alternative protein. This protein was identified as the T3SS injected EHEC protein $EspF_U$. EspF is an effector, encoded by the LEE pathogenicity island, which causes disruption of the intestinal barrier function and ultimately cell death. Notably, $EspF_U$ is encoded by prophage-U from the O157:H7 strain. Apparently this phage protein functions as a potent enhancer of a normally weak signaling activity of EHEC Tir (Campellone *et al.*, 2004).

This is not an isolated case where a prophage-encoded effector protein is injected by T3SS to mediate disruptions of cell functions in the enterocyte. A prominent effect of EPEC infections on epithelial cells is mediated by Cif, which stands for cycle inhibiting factor. Cif triggers an irreversible cytopathic effect by recruiting focal adhesion plaques for the bacteria on the enterocyte. Mechanistically, Cif promotes the actin cytoskeleton rearrangement leading to the assembly of stress fibers and inhibits the cell cycle G2/M phase transition. The cytostatic effect is not functionally related to the stress fiber buildup but is linked to the maintenance of the cyclin-dependent kinase Cdk1 (a driver for entry into mitosis) in a tyrosine-phosphorylated premitotic state. Cif in turn is encoded on a lambdoid prophage of EPEC strains (Marches

et al., 2003). The authors suspected that prophages will encode further effector proteins in EHEC strains. They did not have to wait long to see the fulfillment of their prediction. A Shiga toxin-encoding prophage from an EHEC strain contained near one end of its genome an *nleA*-like gene (Creuzburg *et al.*, 2005). A similar gene had already been described in an EHEC strain and the mouse pathogen *Citrobacter rodentium*. There it was also part of a prophage genome and likewise encoded in a prophage downstream of the tail fiber gene cluster. NleA stands for non-LEE-encoded effector gene A. This bacterial protein is present in the Golgi apparatus after translocation and leads to colonic hyperplasia characteristic for murine *Citrobacter* infections. Mice infected with an *nleA* mutant EHEC strain survived the otherwise lethal infection (Gruenheid *et al.*, 2004). Another case was EspJ, encoded on the cryptic prophage CP-933U of one of the sequenced O157:H7 strains. EspJ influences the clearance dynamics of the pathogen from the host's intestinal tract. In fact, contrary to what one might expect for a virulence gene, *espJ* exhibits antivirulence properties, a not infrequent annotation in prophages. It may favor host survival and thereby aid in pathogen transmission (Dahan *et al.*, 2005). We might here see the transition from virulence to fitness factors where pathogenicity is traded against the evolutionary success of the pathogen, which is frequently linked to its attenuation.

Recently a systematic approach was taken toward T3SS effectors in the O157:H7 strain Sakai. Bioinformatics suggested 60 putative effectors in this strain, and proteomics approaches confirmed the effector status for 39 EHEC proteins (Tobe *et al.*, 2006). Nine of the lambdoid prophages carried effector genes located just downstream of the tail fiber genes. These genes showed an extreme low GC bias suggesting their recent acquisition. The authors speculated that the major function of the large number of lambdoid phages in EHEC is to provide T3SS effectors. Prophages provide an open enormous potential for the evolution of pathogens, allowing potential effector choice from a large phage metagenome.

A very similar story can be told for *Salmonella enterica*. The enterocyte takeover by *Salmonella* is complicated by the presence of two T3SS: T3SS-1 injects effectors that are involved in colonization and invasion of gut epithelia, while T3SS-2 modifies the vacuole in the macrophage in which *Salmonella* resides later on (Schlumberger and Hardt, 2006). In line with this dual task, *Salmonella* has two pathogenicity islands SPI-1 and SPI-2. Prophages also play an important role in the pathogenicity of *Salmonella*. The repertoire of prophages varies between the different strains of the Typhimurium serovar. The different phages encode distinct as well somewhat overlapping functions leading to complicated relationships. For example, strains cured from

prophage Gifsy-2 show a substantially impaired systemic infection in mice, while strains cured from Gifsy-1 maintained their virulence. The impact of Gifsy-1 became only apparent when it was evaluated in strains lacking Gifsy-2. Bioinformatic searches identified a bewildering array of potential virulence/fitness factors encoded on the Gifsy phage family. This list includes GtgE, which is required for full virulence, but also GvrA, which attenuated the pathogenicity as an anti-virulence gene. GipA is required for *Salmonella* replication in the Peyer's patches of the mouse gut, while SodCI protects the pathogen from the oxidative burst in reticuloendothelial cells. GogB, GtgB, and SspH1 are effectors of T3SS, encoded on Gifsy prophages. There are further phages in *Salmonella* one is *sopE*Φ (Mirold *et al.*, 1999). It encodes SopE, a T3SS-secreted protein that acts as a nucleotide exchange factor for host cellular Rho GTPase. SopE plays an important role for bacterial invasion (for a review, see Brüssow *et al.*, 2004). In summary, *Salmonella* prophage genes have even a deeper reach into the eukaryotic host cell than do the prophages from EPEC strains.

3.11.5. *S. pyogenes*/*S. aureus* Phages: The Assault on the Immune System

This long reach of prophage genes can also be illustrated with other bacterial pathogens. *S. pyogenes* and *S. aureus* colonize body surfaces in a substantial proportion of humans, from where they exploit breaches in the skin or mucosa to initiate sometimes life-threatening invasive infections. Overall both pathogens account for an important share of human bacterial infections. Both pathogens target the immune system with virulence factors that overrun the defense of the host by using two different strategies: exhaustion of the immune system by its over-activation and evasion of the immune system. In both pathogens prophages play a prominent role in pathogenicity. In *S. pyogenes* prophages tend to be multiple and they each encode one or two virulence factors, while in *S. aureus* prophages are less numerous, but some encode a complex array of virulence genes. To illustrate the latter point, I will here shortly discuss the innate immune evasion cluster (IEC) encoded near the conserved 3′-ends of several β-hemolysin-converting *S. aureus* phages (van Wamel *et al.*, 2006). An 8-kb genome segment contains the genes *sea-hol-lytA-sak-chp-scn*. In clinical *S. aureus* isolates, these genes are found in various combinations, but always with up to 90% incidence, underlining their crucial importance for *S. aureus*. In contrast, prophages encoding the cytotoxins PVL and ETA are detected in less than 20% of the clinical *S. aureus* isolates.

The different IEC genes have a common denominator: all encode proteins that interfere with the first line of defense of the immune system against invading pathogens. *sea* encodes the staphylococcal enterotoxin A, which induces on human monocytes a down-modulation of cell-surface chemokine binding sites (Rahimpour *et al.*, 1999). Specifically, the binding of monocyte chemotactic proteins (MCP) and of macrophage inflammatory proteins (MIP) is prevented. These proteins belong to the chemokine family of small proteins that comprise chemoattractant cytokines, which play a pivotal role in the extravasation of leukocytes from the circulation. SEA antagonized the directed migration of monocytes to MIP and MCP. SEA thus subverts an important arm of the innate immune system.

The next phage-encoded virulence gene on the IEC, *sak*, encodes staphylokinase, a potent activator of host plasminogen. SAK is used for thrombolysis in myocardial infarction. However, this is not the function for which *S. aureus* acquired staphylokinase. Staphylokinase binds human proteins called α-defensins (Jin *et al.*, 2004). Defensins are a group of peptides secreted by neutrophils. They bind to phospholipids on the bacterial surface, leading to the permeabilization of the microbial membrane by pore formation. Staphylokinase binds to defensin and thus interferes with its potent killing effect. The next gene *chp* encodes CHIPS (chemotaxis inhibitory protein of *Staphylococcus*). CHIPS binds specifically and with high affinity to the C5a receptor and the formylated peptide receptor on human neutrophils. When *S. aureus* invades the human host, complement is rapidly activated, resulting in the opsonization of the bacteria and the generation of large amounts of C5a. C5a and formylated peptides (side products of bacterial translation) are classical chemoattractants and constitute the first triggers of the innate immune system. CHIPS prevents the binding of these ligands (Postma *et al.*, 2004) and thus inhibits the neutrophil influx from the circulation into the peritoneum. The last gene of IEC, *scn*, encodes SCIN (staphylococcal complement inhibitor) (Rooijakkers *et al.*, 2005). SCIN prevents deposition of the activated complement compound C3b on *S. aureus* and thereby reduces neutrophil killing of bacteria. SCIN interferes with the activation of all three complement pathways. Since it acts very early in the complement pathways, it can prevent the generation of the membrane attack complex and spares the bacterium this form of killing.

Among the most frequent virulence genes encoded by *S. pyogenes* prophages are superantigens and DNase. A very attractive hypothesis (Sumby *et al.*, 2005) postulated that the prophage DNase is directed against NET (neutrophil extracellular traps). NET is, however, also an adequate morphological description. NET are very fragile fibers consisting of DNA strands

decorated with histones and anti-bacterial proteases such as neutrophil elastase (Brinkmann *et al.*, 2004). Neutrophils release NET after activation by IL-8 and LPS. The neutrophil elastase degrades virulence factors of Gram-negative bacteria. The fibrous structure of NET is necessary for efficient killing of bacteria. However, killing of bacteria became negligible when the NET were destroyed by the addition of DNase.

Streptococcal superantigens are members of a growing family of proteins that simultaneously bind major histocompatibility complex (MHC) class II molecules and specific variable regions of T-cell receptors. In contrast to normal antigens presented by MHC II, which activate 0.001–0.0001% of all T cells, superantigens activate up to 20% of all T cells. This results in massive T-cell proliferation and subsequent release of inflammatory cytokines. These factors are thought to cause the high fever and shock or autoimmune sequels in some patients with streptococcal infections. While many streptococcal and staphylococcal prophage virulence factors paralyze components of the immune system, we see here possibly an alternative strategy – immune evasion by over-activation of the immune system.

3.12. OUTLOOK

Future research will tell us whether we have in the discussed examples atypical cases for the role of prophages in bacterial pathogenicity or just the tip of a prophage-encoded pathogenicity iceberg. Whatever the outcome, it is clear that phages have entered into a type of cooperation with pathogenic bacteria that make these bacteria dreaded foes for higher organisms. It remains to be seen whether the current efforts to recruit phages for therapeutic approaches against the same bacterial diseases can bind phages into a coalition with humans against human bacterial pathogens.

ACKNOWLEDGMENT

Part of the ideas from this review stems from discussions with Wolf-Dietrich Hardt during the writing of a joint review in 2004.

REFERENCES

Allison, G. E., Angeles, D., Tran-Dinh, N., and Verma, N. K. (2002). Complete genomic sequence of SfV, a serotype-converting temperate bacteriophage of *Shigella flexneri*. *J Bacteriol*, **184**, 1974–87.

Baba, T., Takeuchi, F., Kuroda, M., *et al.* (2002). Genome and virulence determinants of high virulence community-acquired MRSA. *Lancet*, **359**, 1819–27.

Barksdale, L., and Arden, S. B. (1974). Persisting bacteriophage infections, lysogeny, and phage conversions. *Annu Rev Microbiol*, **28**, 265–99.

Barondess, J. J., and Beckwith, J. (1990). A bacterial virulence determinant encoded by lysogenic coliphage lambda. *Nature*, **346**, 871–4.

Bensing, B. A., Siboo, I. R., and Sullam, P. M. (2001). Proteins PblA and PblB of *Streptococcus mitis*, which promote binding to human platelets, are encoded within a lysogenic bacteriophage. *Infect Immun*, **69**, 6186–92.

Beres, S. B., Sylva, G. L., Barbina, K. D., *et al.* (2002). Genome sequence of a serotype M3 strain of group A *Streptococcus*: phage-encoded toxins, the high-virulence phenotype, and clone emergence. *Proc Natl Acad Sci USA*, **99**, 10078–83.

Bielaszewska, M., Fell, M., Greune, L., *et al.* (2004). Characterization of cytolethal distending toxin genes and expression in Shiga toxin-producing *Escherichia coli* strains of non-O157 serogroups. *Infect Immun*, **72**, 1812–6.

Bielaszewska, M., Sinha, B., Kuczius, T., and Karch, H. (2005). Cytolethal distending toxin from Shiga toxin-producing *Escherichia coli* O157 causes irreversible G2/M arrest, inhibition of proliferation, and death of human endothelial cells. *Infect Immun*, **73**, 552–62.

Bordenstein, S. R., Marshall, M. L., Fry, A. J., and Wernegreen, J. J. (2006). The tripartite associations between bacteriophage, *Wolbachia*, and arthropods. *PLoS Biol*, **2**, e43.

Boyd, E. F., and Brüssow, H. (2002). Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. *Trends Microbiol*, **10**, 521–9.

Brinkmann, V., Reichard, U., Goosmann, C. B., *et al.* (2004). Neutrophil extracellular traps kill bacteria. *Science*, **303**, 1532–5.

Broudy, T. B., Pancholi, V., and Fischetti, V. A. (2001). Induction of lysogenic bacteriophage and phage-associated toxin from group a streptococci during coculture with human pharyngeal cells. *Infect Immun*, **69**, 1440–3.

Brüssow, H., and Hendrix, R. W. (2002). Phage genomics: small is beautiful. *Cell*, **108**, 13–6.

Brüssow, H., Canchaya, C., and Hardt, W.-D. (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol Rev*, **68**, 560–602.

Bushman, F. (2002). *Lateral DNA transfer*. New York: CSH Laboratory Press.

Campellone, K. G., Robbins, D., and Leong, J. M. (2004). EspFu is a translocated EHEC effector that interacts with Tir and N-WASP and promotes Nck-independent actin assembly. *Dev Cell*, **7**, 217–28.

Canchaya, C., Desiere, F., McShan, W. M., *et al.* (2002). Genome analysis of an inducible prophage and prophage remnants integrated in the *Streptococcus pyogenes* strain SF370. *Virology*, **302**, 245–58.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., and Brüssow, H. (2003). Prophage genomics. *Microbiol Mol Biol Rev*, **67**, 238–76.

Canchaya, C., Fournous, G., and Brüssow, H. (2004). The impact of prophages on bacterial chromosomes. *Mol Microbiol*, **53**, 9–18.

Casjens, S. (2003). Prophages and bacterial genomics: what have we learned so far? *Mol Microbiol*, **49**, 277–300.

Chen, Y., Golding, I., Sawai, S., Guo, L., and Cox, E. C. (2005). Population fitness and the regulation of *Escherichia coli* genes by bacterial viruses. *PLoS Biol*, **3**, e229.

Conway, J. F., Wikoff, W. R., Cheng, N., *et al.* (2001). Virus maturation involving large subunit rotations and local refolding. *Science*, **292**, 744–8.

Creuzburg, K., Recktenwald, J., Kuhle, V., *et al.* (2005). The Shiga toxin 1-converting bacteriophage BP-4795 encodes an NleA-like type III effector protein. *J Bacteriol* **187**, 8494–8.

Dahan, S., Wiles, S., La Ragione, R. M., *et al.* (2005). EspJ is a prophage-carried type III effector protein of attaching and effacing pathogen that modulates infection dynamics. *Infect Immun*, **73**, 679–86.

Dozois, C. M., Daigle, F., and Curtiss III, R. (2003). Identification of pathogen-specific and conserved genes expressed *in vivo* by an avian pathogenic *Escherichia coli* strain. *Proc Natl Acad Sci USA*, **100**, 247–52.

Edlin, G., Lin, L., and Kudnram, R. (1975). Lambda lysogens of *E. coli* reproduce more rapidly than non-lysogens. *Nature*, **255**, 735–7.

Ferretti, J. J., McShan, W. M., Ajdic, D., *et al.* (2001). Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proc Natl Acad Sci USA*, **98**, 4658–63.

Freeman, V. J. (1951). Studies on the virulence of bacteriophage infected strains of *Corynebacterium diphtheriae*. *J Bacteriol*, **61**, 675–88.

Groth, A. C., Olivares, E. C., Thyagarajan, B., and Calos, M. P. (2000). A phage integrase directs efficient site-specific integration in human cells. *Proc Natl Acad Sci USA*, **97**, 5995–6000.

Gruenheid, S., Sekirov, I., Thomas, N. A., *et al.* (2004). Identification and characterization of NleA, a non-LEE-encoded type III translocated virulence factor of enterohaemorrhagic *Escherichia coli* O157:H7. *Mol Microbiol*, **51**, 1233–49.

Huber, K. E., and Waldor, M. K. (2002). Filamentous phage integration requires the host recombinases XerC and XerD. *Nature*, **417**, 656–9.

Hurst, M. R., Glare, T. R., and Jackson, T. A. (2004). Cloning *Serratia entomophila* antifeeding genes – a putative defective prophage active against the grass grub *Costelytra zealandica*. *J Bacteriol*, **186**, 5116–28.

Jiang, S. C., and Paul, J. H. (1998). Gene transfer by transduction in the marine environment. *Appl Environ Microbiol*, **64**, 2780–7.

Jin, T., Bokarewa, M., Foster, T., *et al.* (2004). *Staphylococcus aureus* resists human defensins by production of staphylokinase, a novel bacterial evasion mechanism. *J Immunol*, **172**, 1169–76.

Kanamaru, S., Leiman, P. G., Kostyuchenko, V. A., *et al.* (2002). Structure of the cell-puncturing device of bacteriophage T4. *Nature*, **415**, 553–7.

Kuroda, M., Ohta, T., Uchiyama, I., *et al.* (2001). Whole genome sequencing of methicillin-resistant *Staphylococcus aureus*. *Lancet*, **357**, 1225–40.

Lang, A. S., and Beatty, J. T. (2000). Genetic analysis of a bacterial genetic exchange element: the gene transfer agent of *Rhodobacter capsulatus*. *Proc Natl Acad Sci USA*, **97**, 859–64.

Lara-Tejero, M., and Galan, J. E. (2000). A bacterial toxin that controls cell cycle progression as a deoxyribonuclease I-like protein. *Science*, **290**, 354–7.

Lara-Tejero, M., and Galan, J. E. (2002). Cytolethal distending toxin: limited damage as a strategy to modulate cellular functions. *Trends Microbiol*, **10**, 147–52.

Lawrence, J. G. (1997). Selfish operons and speciation by gene transfer. *Trends Microbiol*, **5**, 355–9.

Lawrence, J. G., and Ochman, H. (1997). Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol*, **44**, 383–97.

Lawrence, J. G., and Ochman, H. (1998). Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci USA*, **95**, 9413–7.

Lawrence, J. G., Hendrix, R. W., and Casjens, S. (2001). Where are the pseudogenes in bacterial genomes? *Trends Microbiol*, **9**, 535–40.

Makino, K., Yokoyama, K., Kubota, Y., *et al.* (1999). Complete nucleotide sequence of the prophage VT2-Sakai carrying the verotoxin 2 genes of the enterohemorrhagic *Escherichia coli* O157:H7 derived from the Sakai outbreak. *Genes Genet Syst*, **74**, 227–39.

Marches, O., Ledger, T. L., Boury, M., *et al.* (2003). Enteropathogenic and eneterohaemorrhagic *Escherichia coli* deliver a novel effector called Cif, which blocks cell cycle G2/M transition. *Mol Microbiol*, **50**, 1553–67.

Matson, E. G., Thompson, M. G., Humphrey, S. B., Zuerner, R. L., and Stanton, T. E. (2005). Identification of genes of VSH-1, a prophage-like gene transfer agent of *Brachyspira hyodysenteriae*. *J Bacteriol*, **187**, 5885–92.

Matz, C., and Kjelleberg, S. (2005). Off the hook – how bacteria survive protozoan grazing. *Trends Microbiol*, **13**, 302–7.

Matz, C., Deines, P., Boenigk, J., *et al.* (2004). Impact of violacein-producing bacteria on survival and feeding of bacteriovorous nano. *Appl Environ Microbiol*, **70**, 1593–9.

McLeod, S. M., Kimsey, H. H., Davies, B. M., and Waldor, M. K. (2005). CTXΦ and *Vibrio cholerae*: exploring a newly recognized type of phage-host cell relationship. *Mol Microbiol*, **57**, 347–56.

Mirold, S., Rabsch, W., Rohde, M., *et al.* (1999). Isolation of a temperate bacteriophage encoding the type III effector protein SopE from an epidemic *Salmonella typhimurium* strain. *Proc Natl Acad Sci USA*, **96**, 9845–50.

Molineux, I. J. (2001). No syringes please, ejection of phage T7 DNA from the virion is enzyme driven. *Mol Microbiol*, **40**, 1–8.

Moran, N. A., Degnan, P. H., Santos, S. R., Dunbar, H. E., and Ochman, H. (2005). The players in a mutualistic symbiosis: insects, bacteria, viruses, and virulence genes. *Proc Natl Acad Sci USA*, **102**, 16919–26.

Nakagawa, I., Kurokawa, K., Yamashita, A., *et al.* (2003). Genome sequence of an M3 strain of *Streptococcus pyogenes* reveals a large-scale genomic rearrangement in invasive strains and new insights into phage evolution. *Genome Res*, **13**, 1042–55.

Nakayama, K., Kanaya, S., Ohnishi, M., Terawaki, Y., and Hayashi, T. (1999). The complete nucleotide sequence of ΦCTX, a cytotoxin-converting phage of *Pseudomonas aeruginosa*: implications for phage evolution and horizontal gene transfer via bacteriophages. *Mol Microbiol*, **31**, 399–419.

Nakayama, K., Takashima, K., Ishihara, H., *et al.* (2000). The R-type pyocin of *Pseudomonas aeruginosa* is related to P2 phage, and the F-type is related to lambda phage. *Mol Microbiol*, **38**, 213–31.

Ohnishi, M., Kurokawa, K., and Hayashi, T. (2001). Diversification of *Escherichia coli* genomes: are bacteriophages the major contributors? *Trends Microbiol*, **9**, 481–5.

Oliver, K. M., Russell, J. A., Moran, N. A., and Hunter, M. S. (2003). Facultative bacterial symbionts in aphids confer resistance to parasitic wasps. *Proc Natl Acad Sci USA*, **100**, 1803–7.

Oliver, K. M., Moran, N. A., and Hunter, M. S. (2005). Variation in resistance to parasitism in aphids is due to symbionts not to host genotype. *Proc Natl Acad Sci USA*, **102**, 12795–800.

Perna, N. T., Plunkett, III, G., Burland, V., *et al.* (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*, **409**, 529–33.

Postma, B., Poppelier, M. J. van Galen, J. C., *et al.* (2004). Chemotaxis inhibitory protein of *Staphylococcus aureus* binds specifically to the C5a and formylated peptide receptor. *J Immunol*, **172**, 6994–7001.

Rahimpour, R., Mitchell, G., Khandaker, M. H., *et al.* (1999). Bacterial superantigens induce down-modulation of CC chemokine responsiveness in human monocytes via an alternative chemokine ligand-independent mechanism. *J Immunol*, **162**, 2299–307.

Robinson, C. M., Sinclair, J. F., Smith, M. J., and O'Brien, A. D. (2006). Shiga toxin of enterohemorrhagic *Escherichia coli* type O157:H7 promotes intestinal colonization. *Proc Natl Acad Sci USA*, **103**, 9667–72.

Rooijakkers, S. H., Ruyken, M., Roos, A., *et al.* (2005). Immune evasion by a staphylococcal complement inhibitor that acts on C3 convertases. *Nat Immunol*, **6**, 920–7.

Ruzin, A., Lindsay, J., and Novick, R. P. (2001). Molecular genetics of SaPI1-a mobile pathogenicity island in *Staphylococcus aureus*. *Mol Microbiol*, **41**, 365–77.

Sakaguchi, Y., Hayashi, T., Kurokawa, K., *et al.* (2005). The genome sequence of *Clostridium botulinum* type C neurotoxin-converting phage and the molecular mechanisms of unstable lysogeny. *Proc Natl Acad Sci USA* **102**: 17472–7.

Saxena, S. K., O'Brien, A. D., and Ackerman, E. J. (1989). Shiga toxin, Shiga-like toxin II variant, and ricin are all single-site RNA *N*-glycosidases of 28 S RNA when microinjected into *Xenopus* oocytes. *J Biol Chem*, **264**, 596–601.

Schlumberger, M. C., and Hardt, W.-D. (2006). *Salmonella* type III secretion effectors: pulling the host cell's strings. *Curr Opin Microbiol*, **9**, 46–54.

Sinkins, S. P., Walker, T., Lynd, A. R., *et al.* (2005). *Wolbachia* variability and host effects on crossing type in *Culex* mosquitoes. *Nature*, **436**, 257–60.

Smoot, L. M., Smoot, J. C., Graham, M. R., *et al.* (2001). Global differential gene expression in response to growth temperature alteration in group A *Streptococcus*. *Proc Natl Acad Sci USA*, **98**, 10416–21.

Smoot, J. C., Barbian, K. D., van Gompel, J. J., *et al.* (2002). Genome sequence and comparative microarray analysis of serotype M18 group A *Streptococcus* strains associated with acute rheumatic fever outbreaks. *Proc Natl Acad Sci USA*, **99**, 4668–73.

Strauch, E., Kaspar, H., Schaudinn, C., *et al.* (2001). Characterization of enterocoloticin, a phage tail-like bacteriocin, and its effect on pathogenic *Yersinia enterocolitica* strains. *Appl Environ Microbiol*, **67**, 5634–42.

Sumby, P., Barbian, K. D., Gardner, D. J., *et al.* (2005). Extracellular deoxyribonuclease made by group A *Streptococcus* assists pathogenesis by enhancing evasion of the innate immune response. *Proc Natl Acad Sci USA*, **102**, 1679–84.

Summer, E. J., Gonzales, C. F., Carlisle, T., *et al.* (2004). *Burkholderia cenocepacia* phage BcepMu and a family of Mu-like phages encoding potential pathogenesis factors. *J Mol Biol*, **340**, 49–65.

Tobe, T., Beatson, S. A., Taniguchi, H., *et al.* (2006). An extensive repertoire of type III secretion effectors in *Escherichia coli* O157 and the role of lambdoid phages in their dissemination. *Proc Natl Acad Sci USA*, **103**, 14941–46.

van Sluys, M. A., de Oliveira, M. C., Monteiro-Vitorello, C. B., *et al.* (2003). Comparative analyses of the complete genome sequences of Pierce's disease and citrus variegated chlorosis strains of *Xylella fastidiosa*. *J Bacteriol*, **185**, 1018–26.

van Wamel, W. J., Rooijakkers, S. H., Ruyken, M., van Kessel, K. P., and van Strijp, J. A. (2006). The innate immune modulators staphylococcal complement inhibitor and chemotaxis inhibitory protein of *Staphylococcus aureus* are located on beta-hemolysin-converting bacteriophages. *J Bacteriol*, **188**, 1310–5.

Wagner, P. L., Acheson, D. W., and Waldor, M. K. (2001a). Human neutrophils and their products induce Shiga toxin production by enterohemorrhagic *Escherichia coli*. *Infect Immun*, **69**, 1934–7.

Wagner P. L., Neely, M. N., Zhang, X. *et al.* (2001b). Role for a phage promoter in Shiga toxin 2 expression from a pathogenic *Escherichia coli* strain. *J Bacteriol*, **183**, 2081–5.

Waldor, M. K., and Friedman, D. I. (2005). Phage regulatory circuits and virulence gene expression. *Curr Opin Microbiol*, **8**, 459–65.

Waldor, M. K., and Mekalanos, J. J. (1996). Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science*, **272**, 1910–4.

Wick, L. M., Qi, W., Lacher, D. W., and Whittam, T. S. (2005). Evolution of genomic content in the stepwise emergence of *Escherichia coli* O157:H7. *J Bacteriol*, **187**, 1783–91.

Wildschutte, H., Wolfe, D. M., Tamewitz, A., and Lawrence, J. G. (2004). Protozoan predation, diversifying selection, and the evolution of antigenic diversity in *Salmonell*a. *Proc Natl Acad Sci USA*, **101**, 10644–9.

Williamson, S. J., Houchin, L. A., McDaniel, L., and Paul, J. H. (2002). Seasonal variation in lysogeny as depicted by prophage induction in Tampa Bay, Florida. *Appl Environ Microbiol*, **68**, 4307–14.

Wommack, K. E., and Colwell, R. R. (2000). Virioplankton: viruses in aquatic ecosystems. *Microbiol Mol Biol Rev*, **64**, 69–114.

Zhang, Z., Greene, B., Thuman-Commike, P. A., *et al.* (2000). Visualization of the maturation transition in bacteriophage P22 by electron cryomicroscopy. *J Mol Biol*, **297**, 615–26.

CHAPTER 4

The Role of Bacteriophages in the Generation and Spread of Bacterial Pathogens

Roger W. Hendrix and Sherwood R. Casjens

4.1. INTRODUCTION

It appears that one of the first things that occurred to Felix d'Herelle when he discovered bacteriophages in 1917 was that these mysterious objects might provide a means of killing bacteria that are pathogenic to humans (Summers, 1999). The still ongoing story of phage therapy, as this approach was called, has been told elsewhere and will not be retold here, but it serves to point out that scientists have been interested in the effects of phages on their hosts since their discovery. d'Herelle believed, and eventually established, that phages are viruses that infect bacteria. However, it was not until the experimental investigations of phages at the dawn of molecular biology in the 1940s and 1950s that it became clear that phages – and for that matter their bacterial hosts – are genetic organisms (Luria and Delbrück, 1943; Hershey and Rotman, 1949; Hershey and Chase, 1952; Stent, 1963), just like fruit flies, corn, and humans, and so could be expected to mutate and evolve.

Although some work was done on the evolution of phages in the 1960s, 1970s, and 1980s, a more detailed understanding of the genetic mechanisms of phage evolution had to wait until the advent of high-throughput DNA sequencing in the 1990s. This is because the genetic history of a phage, while it is to a significant extent encoded in the phage's genome sequence, is largely invisible to our analysis until we can compare that sequence to the genome sequences of other phages. There are currently more than 300 genome sequences for the large dsDNA phages with tails (tailed phages), as well as about 100 for smaller phages of various types, and the results of genome sequence comparisons, described here, have given us a richly detailed picture of some of the mechanisms of phage evolution. At the same time, sequencing of bacterial genomes has reinforced our appreciation of the

ubiquity of phage DNA in bacterial genomes in the form of prophages and of the fact that prophages typically express genes from the prophage state that affect the phenotype of the host cell (e.g., Casjens, 2003). It appears from these and other observations described hereafter that phages have played an important role – possibly a major role – in the evolution of their bacterial hosts.

This chapter will deal primarily with the dsDNA tailed phages (order *Caudovirales*; Fauquet *et al.*, 2005). This is the group for which the most is known about the phages' evolution as well as about the phages' influence on the hosts' phenotype and evolution. However, we will also touch on members of the ssDNA filamentous phages, such as CTXϕ, which forms a prophage in *Vibrio cholerae* and encodes the cholera toxin.

4.2. ABUNDANCE OF PHAGES

Comparative examination of phage genomes implies, as we describe below, that the current genomes are the products of enormous numbers of exceedingly improbable genetic events, most notably non-homologous recombination. The only way this can make sense is if opportunities for carrying out these events are enormously frequent, or if they have been going on for an enormously long time, or, as we now believe is the case, both. The first indications of the size of the global 'phage evolution machine' came from experiments to enumerate tailed phages in an environmental sample directly, by electron microscopy, rather than by asking how many plaques are produced on a particular laboratory strain of a particular species of bacteria. The startling result was that there are up to 10^7 tailed-phage particles in a milliliter of Norwegian fjord water (Bergh *et al.*, 1989), and comparable and higher measurements have now been made in many aquatic, marine and terrestrial environmental locations (Angly *et al.*, 2006; Hambly and Suttle, 2005; Danovaro and Serresi, 2000; Hewson *et al.*, 2001; Suttle, 2005). Based on these numbers, we have estimated that there are on the order of 10^{31} individual phage particles in the biosphere (Whitman *et al.*, 1998; Hendrix, 2003). This is a truly astronomical number in that if the phage particles were laid end-to-end they would stretch into space for 200 million light years. Ecological measurements suggest that this entire population turns over every few days, from which we calculate that on a global scale there must be roughly 10^{24} productive infections per second to replenish the population as it is depleted by infection of new hosts, physical damage, and predation by protozoa that eat viruses (Garza and Suttle, 1998; Weinbauer *et al.*, 1993). Each of these infections provides an opportunity for genetic change in the phage's genome,

through point mutation or homologous or non-homologous recombination. It is such changes that generate the genetic diversity in the population that is the raw material of evolution.

4.3. MECHANISMS OF PHAGE EVOLUTION

4.3.1. Point Mutations

Like any other genes, phage genes show evidence of incremental change through point mutation. Contrary to what is seen in cellular organisms, there are no genes in the genomes of tailed phages analogous to the rRNA genes of cellular genomes that can be connected into a single family spanning all known genome sequences, using either the nucleotide sequences or the encoded protein sequences (e.g., Casjens, 2005). Perhaps the most obvious candidate for such an 'rRNA-like' gene is the gene encoding the major capsid (or coat) protein. If we choose one of these proteins and search for similar sequences in the sequence databases (using, for example, the PSI-BLAST algorithm; Altschul *et al.*, 1997) we typically find a large collection of sequences that vary from 100% amino acid identity to levels of amino acid identity on the very edge of detectability (at or below about 12% identity). Capsid proteins fall into a limited number of such non-overlapping lineages that can be inferred from sequence comparisons. Recent structural studies are beginning to provide evidence for a common polypeptide fold in coat proteins which is shared across all of the tailed-phage sequence-defined lineages (and even by the major capsid subunit of herpes viruses! – Baker *et al.*, 2005), arguing for ancestral connections between even the different sequence-defined lineages (Wikoff *et al.*, 2000; Jiang *et al.*, 2003; Fokine *et al.*, 2005; Morais *et al.*, 2005). A picture is emerging in which all of the coat proteins are most likely homologous (i.e., they share common ancestry) and they have all preserved the ancestral protein fold, but the individual amino acid sequences have accumulated point mutations and diverged to such an extent that most randomly chosen pairs of capsid proteins are not detectably related in sequence.

Similar pictures of common ancestry and extreme sequence divergence are becoming apparent for other phage genes, including a few other structural protein genes such as the head portal and terminase (packaging DNA translocase) genes that are present in all of the tailed phages (Casjens, 2003), as well as some of the genes encoding proteins present in only a subset of the tailed phages, like integrases, DNA and RNA polymerases, repressors, etc. The high levels of divergence through point mutation bespeak a very

long history for these phages. Unfortunately, it is not possible to calibrate that time scale with any degree of accuracy because we do not know how fast the 'mutational clock' runs for phages in the natural environment. With that caveat, however, the degree of sequence divergence observed in phage proteins, together with the evidence cited earlier that the tailed-phage and herpesvirus capsid proteins may have common ancestry, favors the view that the history of these phages spans billions of years rather than tens or hundreds of millions of years. If so, the pervasive influence of phages on the evolution and phenotypes of their hosts that is evident in contemporary biology (see later discussion) may have been in effect since near the beginnings of cellular life (Koonin *et al.*, 2006; Forterre, 2006).

4.3.2. Moron Addition

Known genomes for tailed phages range in size from 19 kb to 500 kb. Any plausible model for phage evolution must incorporate a mechanism for increasing the size of the genome, and in fact there is ample evidence for additions of genes or groups of genes to a phage genome. We consider first a type of genetic element associated with such additions, named a 'moron' (for 'unit of more DNA', Hendrix *et al.*, 2000; Juhala *et al.*, 2000). A moron is typically a single protein-coding gene flanked by a transcription promoter and a transcription terminator. It is identified as a recent addition to the genome by the fact that it is found inserted between two genes that are adjacent in a closely related phage. In favorable cases, the moron DNA has a different G+C content from the flanking DNA, suggesting a foreign source and therefore arguing that it is indeed a recent addition to the genome rather than evidence of a recent deletion from the comparison phage. It is not clear what the source of morons is, nor is it clear by what biochemical mechanism they insert into a recipient genome. However, in several cases there is something known about the function of morons, and in most of these cases they are functions that are expected to be beneficial to the host cell when they are expressed from a prophage. Examples include *cor*, a gene in *Escherichia coli* phages ϕ80, HK022, N15, ES18, and mEp167 that encodes an outer membrane protein that blocks adsorption by a group of phages that use the same receptor (Uc-Mass *et al.*, 2004), or the *sod* gene of *Salmonella* phage Gifsy-2, which encodes a superoxide dismutase (Figueroa-Bossi and Bossi, 1999). We speculate that many moron genes benefit the phage indirectly, by providing a direct benefit to the host cell from the prophage and therefore a disincentive for the cell to delete the prophage. The implications of this for bacterial evolution are discussed in more detail later.

In principle morons can also provide a selective benefit to the phage by enhancing its lytic growth; such a role has been suggested, for example, for the *D* gene of *E. coli* phage λ (Sternberg and Weisberg, 1977; Morgan *et al.*, 2002) and the *dec* gene of *Salmonella* phage L (Gilcrease *et al.*, 2005; Tang *et al.*, 2006), both of which encode accessory proteins that add to the capsid during its maturation and stabilize it. More generally, a case can be made that phages have acquired a substantial fraction of their genes through a process of 'moron accretion', in which the phage genome is continually bombarded (on an evolutionary time scale) by morons that insert into the sequence (Hendrix *et al.*, 2000). Morons that are selectively neutral or detrimental will be lost from the population but the (presumably rare) moron that provides a selective benefit will tend to be retained. Over time, mutations that serve to integrate the expression of the moron into the regulatory circuitry of the phage will be selected and the moron will gradually look less like an interloper and more like a bona fide phage gene.

4.3.3. Non-homologous Recombination

The most striking conclusion from genome sequence comparisons is that the tailed phages are genetic mosaics with respect to each other. If two genome sequences are lined up with each other, and assuming they are sufficiently closely related to have some sequence similarity, that similarity will be found in patches (e.g., Botstein, 1980; Pedulla *et al.*, 2003; Casjens, 2005; Hendrix, 2003). We might see, for example, two genes that align between the two phages at a high level of identity, followed by an abrupt transition to a lack of any detectable similarity, either because the adjacent genes have diverged extensively as described earlier or, as often, because the adjacent genes are from different, non-homologous gene families. The explanation for such an observation is that there has been a non-homologous recombination event – that is, a recombination between different sequences, producing a novel juxtaposition of sequences – occurring in the ancestry of one of the two phages. For a given pair of phages there may be as many as 50 such novel joints visible in a sequence comparison, serving as fossils of past genetic events. This is in fact an underestimate since novel joints that were formed in the shared ancestry of both phages will not be visible, nor will novel joints for which both halves were in the ancestry of only one of the two sequences being compared.

The sites of non-homologous recombination identified in this way are not randomly distributed across the genome sequence. Rather, in most cases they are located near gene boundaries. In many cases the mosaic modules

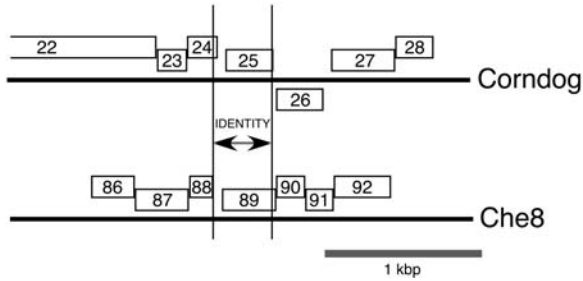


Figure 4.1. Comparison of portions of the gene maps of mycobacteriophages Corndog and Che8, surrounding the regions where they share 378 bp of 100% nucleotide sequence identity. The region of identity, indicated by vertical lines and a double arrow, is inferred to have moved horizontally from one lineage to the other in the recent ancestry of these phages. The flanking regions show no detectable sequence similarity. Although the sequences are aligned on the identity, they come from different parts of the two genomes, as suggested by the gene numbers. All genes shown are transcribed left to right except for Corndog *26*.

defined by flanking novel joints are single genes; sometimes they are groups of genes with related functions such as the head genes. Some novel joints fall within coding regions, but in these cases their location typically corresponds to a domain boundary of the encoded protein. These are all locations where we can imagine that two different sequences could be spliced together without damaging the function of the new genome produced as a result. The current view, generated from the examination of phage sequences, is that non-homologous recombination occurs rarely but promiscuously and quasi-randomly across the genome, producing novel joints within coding regions as well as at gene boundaries, and in-register and out-of-register with respect to genome organization. The vast majority of such recombinants will be functionally compromised, and only those that function as well as or better than their parents are expected to survive and be available for us to determine their genomic sequence. Evidence for this view comes largely from examining actual examples of novel joints where we see that the recombination has not always made a tidy junction but only one that is close enough to 'tidy' to appear non-destructive of function (Hendrix, 2002). An example of one such recombinant is shown in Figure 4.1. The two phages compared here, mycobacteriophages Che8 and Corndog, are very dissimilar in sequence, with only patches of low similarity scattered across their genomes, except for one 378 bp stretch of 100% nucleotide sequence identity between the two phages. This observation implies that there has been a transfer of sequence

by non-homologous recombination between the recent ancestors of these two phages. The fact that there have not been any point mutations in either lineage since the transfer argues that the sequences we see are most likely the result of the primary recombination event, unobscured by any additional 'tidying' deletions or insertions. Interestingly, these two phages were isolated in different parts of the world (Chenai, India, and Pittsburgh, USA), implying that there may be no barriers to the global spread of phages on an evolutionary time scale.

It may seem implausible that phages could have produced the huge variety of functional mosaic genomes that we observe by such an undirected mechanism, using intrinsically highly improbable events that almost always produce non-functional genomic junk. However, this is where our knowledge of the size and dynamic nature of the global phage population can help. The roughly 10^{24} productive phage infections per second on a global scale (mentioned earlier) have, in almost all cases, the potential to produce recombinants with resident prophage DNA or, less frequently, with DNA of a co-infecting phage (Casjens, 2003). Another factor that would seem to increase the probability of the success of such events is the tendency to keep the order of gene functions constant, especially (but not limited to) within phage types (such as the lambdoid phages), even when there is little sequence similarity between some members of the group (Casjens *et al.*, 1992). If we assume (somewhat arbitrarily) that a non-homologous recombination event occurs in one in 10^9 of these infections, and if we assume (again somewhat arbitrarily) that only one in 10^9 of these recombinants is sufficiently functional to survive the scrutiny of natural selection, then there will still be 10^6 novel joints produced per second that have the potential to turn up in the genomes we sequence.

4.3.4. Homologous Recombination

In laboratory experiments, homologous recombination is several orders of magnitude more frequent than non-homologous recombination, and we presume the same must be true in natural settings. Homologous recombination does not create novel joints in the sequence, and so the direct effect is not visible in sequence comparisons. However, it does have the potential to reassort the novel combinations of sequence created by non-homologous recombination into new genomic groupings, and it also may serve to distribute the novel combinations rapidly through the global population of phages. As the mosaicism of phages was first revealed by DNA heteroduplex experiments in the 1960s and 1970s, it was suggested that there might be some similar

feature of gene boundaries that could be utilized by homologous recombination to generate new novel joint combinations (Botstein, 1980; Susskind and Botstein, 1978). Such sequences, now called 'boundary sequences' (Casjens *et al.*, 2004b), may be present in some cases and may be partially responsible for new combinations of novel joints; however, homologous recombination between homologous genes must be a major contributor as well.

4.3.5. Prophage Role in Phage Recombination

As mentioned earlier, prophages are also very abundant, with more than one prophage per genome on average among the bacterial genomes analyzed, and a significant fraction of phage genes present on Earth are in prophages (Canchaya *et al.*, 2003b; Casjens, 2003; Fouts, 2006; Mehta *et al.*, 2004; Bose and Barber, 2006). Prophages are thought to have an important role in phage evolution because they provide a partner for recombination with any phage infecting a lysogen. In addition, prophages have the potential to recombine with other prophages in the same genome. A dramatic example of this is seen in a comparison of the pathogenic *E. coli* strains EDL933 and Sakai (Casjens, 2003). These strains each have 18 to 20 prophages (not all intact) in their genomes, and a number of these are in corresponding positions in the two bacterial genomes. The central portions of at least four of these prophages have swapped positions in one cell in comparison to their positions in the other cell, apparently by homologous or near-homologous recombination between the prophages. These new combinations of genes or new genomic arrangements produced as a result of these recombinations are available for recombination with infecting phages in addition to direct phage generation if the prophage is fully functional.

4.4. NATURE OF THE PHAGE POPULATION

Despite the increased number of phage genome sequences available in recent years, the total number is still a minuscule fraction of the global phage population. Worse, the available sequences are very strongly biased toward phages that infect a very restricted group of bacterial hosts, primarily those such as *Escherichia, Salmonella,* or *Bacillus* that have been favored model systems in molecular biology, or those like *Lactobacillus* or *Mycobacterium* that have commercial or medical interest. With this background as a caveat, we can ask what can be deduced about the genetic structure of the phage population as a whole. First, it is clear that the population is extremely diverse; to our knowledge no two independently isolated phages have identical genome

sequences, and most pairs of phages chosen from the sequenced population have no recognizable shared sequence; identical phage *genes* have been isolated independently, however (Breitbart *et al.*, 2004b). When a new phage is sequenced, particularly if it is the first phage to be sequenced from a new group of hosts, it is typical for only about one-third of the predicted proteins encoded by the phage to make detectable matches to the sequence databases. Second, phages do cluster into types to some degree. For phages of *E. coli* and phylogenetically related hosts, phages resembling the classically studied types such as λ, T4, and T7 are found repeatedly, but even in these groups the resemblance may be only in genome size and general gene organization and content. Furthermore, some groups of phages, typified by *E. coli* phage N15, have turned out to be apparent hybrids relative to previously recognized types. When we move to more distantly related hosts such as *Mycobacterium* and *Streptomyces,* we also find many genomes clustering into types, but these do not fit into any of the genome types of *E. coli* phages. Despite the impression of extreme diversity, however, it must also be said that all known tailed phages are vastly more similar to each other than would be expected if their genes had been assembled at random from the global pool of phage genes, in terms of each phage having a suite of genes suitable for its particular lifestyle, having them clustered on the genome according to function, and having them arranged to facilitate an orderly temporal expression that benefits evolutionary success of the phage (Hatfull *et al.*, 2006). In sum, there appears to be an extremely diverse population that is nevertheless not distributed randomly through genome space but rather clustered loosely around a large number of fundamentally similar solutions to the problem of how to be a successful bacteriophage.

A complementary view of the population structure comes from metagenomic studies, in which all the phages from an environmental sample are collected and sequenced in bulk (Edwards and Rohwer, 2005). In principle, and also probably in practice, this approach gives a more unbiased picture of the kinds and frequencies of sequences in the population. Its chief drawback is that all sequences in the population are mixed together, so information about how genes are grouped and arranged in individual genomes is lost. A striking conclusion from metagenomic studies is that the diversity of types of phages and the diversity of sequences are greater than has been measured for any other group of biological organisms (Breitbart *et al.*, 2004a). When this information is combined with the evidence for 'rampant' horizontal exchange of genes across the phage population through non-homologous recombination, we get a picture in which the global phage population is the repository of an unprecedented diversity of genetic sequences and in which

all phages in the population have access (directly or indirectly) to any of those sequences. Considering that phage genes can move with relative ease into bacterial genomes, this huge reservoir of genetic diversity is also available to the phages' hosts.

When translations of metagenomic sequences are compared to the protein sequence databases, as described earlier for individual genomes, the results are similar: matches are found for about one-third of the sequences. This lends some support to the hope that the tiny minority of phages that can be grown in the lab and subjected to whole genome sequencing are representative of the phage population as a whole. More definitive tests of this idea will most likely require radical improvements in sequencing technology.

4.5. PHAGE EFFECTS ON HOST EVOLUTION

4.5.1. Phage Lytic Growth and Bacterial Evolution

Bacteria live in a world in which there are about 10 tailed-phage particles for every bacterial cell. Although there are many ways in which phage infection can be beneficial to a bacterial lineage on an evolutionary time scale (see later discussion), a more immediate problem for bacteria is how to avoid infection by the phages in its environment, since most such encounters will simply kill the cell. Thus, there is a strong selective pressure for bacterial mutants that are resistant to infection by the phages that are killing other members of their population. Once such phage-resistant mutants become abundant in the population there will be a selective advantage created for phage mutants that are able to overcome the block to infection and once again infect successfully. Such an 'arms race' between virus and host has presumably been happening since the origins of phages.

The simplest type of defensive mutation in bacteria, and the first to be found in the early genetic experiments with phages (Luria and Delbrück, 1943; Weitz *et al.*, 2005), is a mutation in the cell surface receptor that prevents infection by preventing binding of the phage to the host cell. However, mutations that sabotage phage production can occur at any point in the growth cycle where the phage depends on interactions with host cell components, and many such examples are known (e.g., Georgopoulos *et al.*, 1973). A particularly colorful example is seen in the *prrC* gene found in some strains of *E. coli*. When such a cell detects an infection by a phage in the T4-like group it activates a nuclease that inactivates an essential $tRNA_{Lys}$ by cleaving it in the anticodon (Penner *et al.*, 1995). This cellular self-immolation succeeds in aborting the infection for some phages in this group and thereby

presumably benefits the cell's siblings in the population by preventing the production of phages that could infect them. However, T4 itself counters this cellular defense by elaborating an RNA ligase that repairs the inactivated tRNA.

4.5.2. Phages as Vectors for Horizontal Transmission of Bacterial Genes

4.5.2.1. Transduction of Genetic Information

The advent of whole bacterial genome sequencing has produced overwhelming evidence for widespread and ongoing horizontal exchange of genetic information between different bacterial species (summarized by Ochman *et al.*, 2005). Many phages have the ability to carry out such DNA movement, which is called transduction when it is phage-mediated. The several classes of phages that are able to perform transduction are also ancient (Casjens *et al.*, 2005), suggesting that phages have been moving their hosts' DNA around nearly as long as phages have existed and essentially since bacteria came into existence (as discussed earlier). Transduction can occur by several different mechanisms, and we do not review all these here, but the simplest and likely most common mechanism is a simple DNA packaging error in which a phage genome sized fragment (typically in the 35–100 kb size range) of the chromosome of the infected cell is packaged instead of phage DNA. The resulting virions are functional and can deliver their DNA into another bacterial cell. Since there are no phage genes present, such an infection may not harm the cell and the DNA injected is free to recombine into the target cell's genome by any means available. Unfortunately, transduction typically leaves no phage-specific tracks in the target genome, so it is very difficult to determine the mechanism by which past horizontal transfer events occurred. Nonetheless, given the number of phage infections per second on Earth (mentioned earlier), the fraction of tailed phages that can perform transduction (10% is a reasonable and conservative estimate; 29 of 114 randomly chosen phages and prophages have packaging enzyme amino acid sequences associated with generalized transduction (Casjens *et al.*, 2005), and the fraction of such phages that carry host DNA (about 0.01% for *Salmonella* phage P22; Ebel-Tsipis *et al.*, 1972) suggest that something like 10^{19} tailed-phage virions deliver a payload of purely bacterial DNA every second (see also Jiang and Paul, 1998). Many phages are naturally able to infect several related species of host, and one can imagine rarer accidental delivery into less related bacteria. It seems an inescapable conclusion that phages must have been, over the eons, major

players in the diversification of bacteria by horizontal transfer of genetic information.

Some bacteria have a special form of 'transduction'. The genome of these bacteria, which include *Rhodobacter capsulatus* and *Brachyspira hyodysenteriae*, encode tailed-phage-like devices called 'gene transfer agents' in which under certain circumstances expression of the genes is turned on and the encoded phage-like proteins build virions that contain only random fragments (in these cases 4–8 kb in length) of the bacterial genome. These phage-like particles can, like other generalized transducing particles, inject their DNA into other members of the species from which they came (reviewed in Lang and Beatty, 2007; Casjens, 2003). It can be argued from their universal presence in the species and the way in which they are controlled that these are not accidentally partially functional defective prophages. If this is true, then gene transfer agents must have evolutionary value to these bacterial species, but we do not yet know exactly what that value is.

4.5.2.2. Phages and Diversity within Bacterial Gene Families

Phages, lytic and temperate, often carry homologues of important host genes that presumably perform better for the purposes of phage replication. These genes are under different evolutionary pressures than are the host genes and so generate additional evolutionary diversity in those functions. For example, many phages carry genes whose encoded proteins have functions in nucleic acid metabolism, such as polymerases, helicases, nucleases, DNA modification enzymes, recombination machinery, and single-strand DNA binding. Three other examples are as follows: (1) Recent sequencing of the genomes of tailed phages that infect photosynthetic cyanobacteria revealed that they often carry genes which encode proteins that by homology are predicted to participate in the photosynthetic process of the host bacterium (Lindell *et al.*, 2004, 2005; Mann *et al.*, 2003; Millard *et al.*, 2004; Weigele *et al.*, 2007). (2) Similarly, *Bacillus subtilis* phages SP10 and PMB12 carry functions that can compensate for some early host sporulation defects (Silver-Mysliwiec and Bramucci, 1990), and it seems likely that these functions might be related to the proteins of the host sporulation machinery. (3) Another example is a phage tail protein that stimulates stationary phase mycobacteria to grow (Piuri and Hatfull, 2006). It seems likely that at times such new phage-generated diversity would fortuitously give a function that the host could appropriate for its own benefit. Zeidner *et al.* (2005) and Sullivan *et al.* (2006) have presented evidence and argued that this is indeed the case with the phage-borne photosynthesis genes.

4.6. PROPHAGES AND BACTERIAL EVOLUTION

4.6.1. Prophages and Bacterial Genome Structure

Although the fact that many if not most bacterial genomes harbor prophages has been known anecdotally for about five decades, the analysis of the complete sequences of bacterial genomes has brought this home spectacularly (Canchaya *et al.*, 2003b; Casjens, 2003). Some bacteria may carry more than 15 prophages. These prophages can have many different influences on their hosts and so will influence their evolution. Such prophages may come and go in a bacterial lineage rapidly on an evolutionary scale, but when a mutation occurs that locks them into the bacterial chromosome (for example, a deletion of the integration site on one side of the prophage) it may take millions of years for the now useless phage DNA to be removed by natural bacterial mutational processes (Casjens, 2003; Lawrence *et al.*, 2001). In either case, the presence of prophage DNA can have important consequences on the evolution of its host. Various aspects of this interaction have been reviewed (Boyd and Brüssow, 2002; Brüssow *et al.*, 2004; Canchaya *et al.*, 2003a, 2004; Casjens, 2003; Casjens and Hendrix, 2005; Wagner and Waldor, 2002), so we focus here on only a few examples to illustrate points and do not attempt to review the field comprehensively.

Given the abundance of phage virions, the frequency of phage infections of bacteria in Earth's biosphere (as discussed earlier), and the frequency with which prophages are found in bacterial genomes, one can wonder why bacterial genomes do not become hugely bloated with integrated phage DNA. It seems likely that bacteria have achieved an optimum rate for the occurrence of 'non-homologous' deletions, a rate that should respond to evolutionary pressures, compromising between too-frequent accidental deletion of their own essential genes and the inability to eventually remove integrating prophages and transposons and other genetic parasites (Lawrence *et al.*, 2001). Thus prophages, along with other integrating DNA elements, must have contributed to the nature of bacterial chromosomes by forcing bacteria to allow random (or nearly so) deletions to occur at a significant frequency. This in turn may have also contributed to the rather rapid removal of genes that have become unnecessary when bacteria colonize a new niche.

Phages have also contributed to the physical structure of bacterial chromosomes in another way. When two related phages integrate at different locations in a bacterial chromosome, they can carry similar sequences that

can serve as sites for homologous recombination. Such recombination events can either invert or remove the DNA between the prophages, depending upon the relative orientation of the prophages. Both these events could be detrimental and thus phage and bacterial mechanisms to avoid them should be selected for. We do not yet know what these might be, but it has been observed that the lambdoid phages that infect the γ-purple bacteria nearly always integrate in one orientation relative to the chromosome in one replichore and the opposite direction in the other replichore (Campbell, 2002). Thus, if recombination is more likely between replichores than within them, such events between prophages would cause inversions rather than deletions, and inversions seem less likely to be detrimental. Analysis of bacterial genome sequences has shown that closely related species or different individuals within species often carry inversions relative to one another that could have been caused by this type of reciprocal recombination event between the two replichores. In most cases too much time has passed to identify the exact points of recombination; however, in some specific cases such events can be convincingly ascribed to recombination between two related integrated prophages in, for example, *Streptococcus pyogenes* (Nakagawa *et al.*, 2003) and *E. coli* (Iguchi *et al.*, 2006).

Phages usually avoid damaging the host genetic information upon integration; however, some phages, most notably *E. coli* phage Mu, integrate by a transposition mechanism with little or no target specificity, so they may integrate into and thus disrupt a host gene. Presumably, the rare occasions when such phages integrate into essential genes are not detrimental enough to the phage to force a lifestyle change. On the other hand, most temperate phages integrate into specific chromosomal sequences. Some of these are between genes and so might not be expected to affect the host's genetic information; however, some integrate into protein coding genes and many integrate into tRNA or tmRNA genes, and when they do this they replace the separated 3′-part of the gene with a phage-encoded replacement, so that a functional gene is regenerated at one end of the integrated prophage (Williams, 2002). The frequent selection of stable RNA genes for integration targets may reflect their slow evolution (they are difficult to change without altering function) and hence the host cannot easily mutate the integration site to avoid the prophage. The phage-carried replacement parts of these genes are not always identical to the sequences they replace, and so this process could allow bacteria to search more 'sequence space' for the affected genes via this process than they could without it.

4.6.2. Expression of Prophage Genes, Lysogenic Conversion, and Bacterial Pathogenicity

4.6.2.1. Prophage Gene Expression

Most genes of integrated prophages are not expressed, and these are genes that are only useful to the phage during lytic growth. However, temperate phages must express a repressor that keeps those genes off and so must have at least one gene (the repressor gene) that is turned on in the prophage state. But it is not this simple; for example phage λ of *E. coli* encodes five other proteins that are made from the prophage (RexA, RexB, SieB, Lom, and Bor); the first three of these appear to have functions that protect the lysogen from infection by other phages (called 'superinfection exclusion') and the last two affect the bacterium's interaction with mammalian hosts (see Hendrix and Casjens, 2006, for details of their functions). We note that it has been reported that λ lysogens grow more rapidly in culture than isogenic non-lysogens in the absence of either of these factors (Edlin *et al.*, 1975), so we do not yet understand all of the prophage-host interactions. The genes encoding these proteins are called 'lysogenic conversion' genes, because their presence converts or modifies the properties of the bacterium harboring the prophage. These themes have now been reiterated many times in many variations with many prophages. Prophages often express proteins that protect the lysogen from attack by other particular (not all) phages, and they also express proteins that affect the interaction of a bacterial pathogen with its eukaryotic host. Clearly, the expression of such phage genes will affect many aspects of the evolution of the host in ways that are difficult to predict. We next discuss a few specific examples (out of many) of lysogenic conversion to give the reader the flavor of the state of our knowledge in this arena.

4.6.2.2. The Prophage Repressor

Like other prophage repressors, the phage λ repressor binds the operators for the early phage operons, so when present it blocks transcription and the expression cascade of lytic gene functions. Thus, in addition to keeping the prophage genes off, repressor will also bind to infecting phages' early operators, if they are similar enough to its own, blocking their lytic development. This phenomenon is called 'immunity' and is different from superinfection exclusion (discussed earlier). In addition to this role, phage λ repressor has recently been shown to specifically partially repress a host gene, *pckA*, which encodes phosphoenolpyruvate carboxykinase, an enzyme

required for the conversion of oxaloacetate and adenosine triphosphate (ATP) into phosphoenolpyruvate, a function that is important in the utilization of carbon sources that feed into oxaloacetate and for gluconeogenesis (Chen *et al.*, 2005). It is not yet known why *E. coli* finds it advantageous to turn down expression of this gene when it becomes a lysogen, but the striking observation that the operator region for the *pckA* gene also has binding sites for the repressors of temperate *E. coli* phages 21, H-19B and 434 (which have different DNA-binding specificities from λ repressor) strongly suggests that evolutionary pressure in this direction has existed for a long time (long enough to have 'collected' the various operators). It might also suggest that different lambdoid prophages have come and gone in this bacterial lineage, since these operators would seem to be of no use in the absence of the phage repressor.

4.6.3. Prophages and Bacterial Virulence

4.6.3.1. Phage Resistance and the Bacterial Surface

Prophage lysogenic conversion genes often protect the lysogen from super-infection by other specific phages. These super-infection exclusion mechanisms are quite varied, and we will mention only one type here, genes encoding proteins that alter the surface receptors to prevent adsorption or injection of DNA from the excluded phage virions. Even these are known to occur by several different mechanisms.

As mentioned previously, a number of *E. coli* and *Salmonella* phages adsorb to the outer membrane protein FhuA, a component of the ferrichrome (an iron siderophore) transport system. The temperate phages that use this receptor also carry a gene, called *cor*, which encodes a short protein that inactivates the FhuA protein for both phage adsorption and ferrichrome transport (Killmann *et al.*, 2001; Matsumoto *et al.*, 1985; Uc-Mass *et al.*, 2004). It is not known whether it acts directly on the FhuA protein or the TonB mediated system that activates it.

Shigella is a major cause of bacillary dysentery, and like many pathogens it uses a type III secretion system (discussed later) to inject effector proteins into host cells to achieve its purposes. This leads to an intense inflammatory response in the gut mucosa of the host, but the bacterial cell is at least partially protected from this response by the polysaccharide portion of the lipopolysaccharide (called O-antigen) on its surface. *Shigella flexneri* is the most common *Shigella*, and it is present with at least 13 different O-antigens (or serotypes), 12 of which are modifications of the basic 'Y type' polysaccharide that contains

a repeating tetrasaccharide *N*-acetylglucosamine-rhamnose-rhamnose-rhamnose. These modifications have at least two consequences: They can protect against phages that utilize the O-antigen as a receptor, and they are considered to be important virulence factors because the host has to mount a specific immune response to each serotype to be protected from it. Temperate phage lysogenic conversion genes are responsible for most if not all of these modifications (Allison and Verma, 2000), and these phage-encoded modifying enzymes, which glucosylate or acetylate the O-antigen at different positions, have been characterized from phages SfII, SfV, Sf6, and SfX (Allison and Verma, 2000; Korres and Verma, 2006; Markine-Goriaynoff *et al.*, 2004). These modifications may also have yet another important effect. The O-antigen chains in some cases may extend beyond the type III secretion needle and sterically impede contact with the target cell. In at least the case with the phage SfV-encoded enzyme, the glucosylation appears to cause the O-antigen to adopt a more compact conformation, thus exposing the needle (West *et al.*, 2005). O-antigen modification is not restricted to *Shigella* prophages but has also, for example, been found in *Pseudomonas* phages (Newton *et al.*, 2001) and *Salmonella* phage P22 where expression of the glucosyltransferase is the subject of an on/off phase variation (Fukazawa and Hartman, 1964).

4.6.3.2. Shiga-like Toxins

The first bacterial virulence factor known to be carried by a prophage was diphtheria toxin (Freeman, 1951; Uchida *et al.*, 1971), and the *E. coli* O157:H7 Shiga toxin gene has also been known to be phage-borne for some time, but the high frequency of this type of relationship was not realized until the advent of complete bacterial genome sequencing. We now know that it is very common for toxin and other virulence factor genes to be carried on phage genomes. One of the best studied is the *E. coli* O157:H7 that can cause severe, life-threatening hemorrhagic diarrhea in humans (reviewed by Herold *et al.*, 2004). Upon its release from the bacteria the toxin is taken up by intestinal cells where it enzymatically cleaves the rRNA, thereby killing the cell. The phage-borne Shiga toxin genes are carried as an apparent moron between the late promoter and the lysis genes of integrated lambdoid prophages of approximately 60 kb, known as '*stx*-converting phages'. The control of the expression of this toxin is complex; it is expressed at considerably higher levels after the prophage is induced to grow lytically, and the toxin may rely on phage-mediated cell lysis to be released from the bacterium (Wagner *et al.*, 2002).

4.6.3.3. Cholera Toxin and ssDNA Filamentous Phages

Phage-borne bacterial virulence genes are not limited to the Caudovirales (tailed phages). Cholera toxin, a critically important factor in cholera epidemic causative *Vibrio cholerae*, is encoded by two genes present on the small (~7 kb) temperate filamentous phage CTXΦ that integrates at a particular site in the bacterial chromosome (Waldor and Mekalanos, 1996). This toxin is largely responsible for the life-threatening diarrhea that accompanies cholera. The bacterium secretes the toxin as it colonizes the mucosa of the human small intestine. The toxin is taken up by intestinal cells and ADP-ribosylates the signaling G proteins, which eventually results in massive efflux of water and salts from the cells and severe diarrhea (De Haan and Hirst, 2004). The toxin genes are thought to be expressed sufficiently from the uninduced prophage to exert their pathogenic effect (Quinones *et al.*, 2006a, b), so, unlike Shiga-like toxin production (discussed earlier), prophage induction does not seem to play a role. Cholera toxin is secreted from the bacterial cell via a type II secretion system (Sandkvist, 2001). Similar prophages of filamentous phages have been found recently in the genomes of *E. coli* K1 strains (prophage CUS-1), invasive *Neisseria meningitidis* strains (prophage MDA), and *Yersinia pestis* (prophage YpfΦ) that have been associated with, respectively, virulence in extra-intestinal infections, cerebrospinal meningitis, and the biovar of the third (current) plague pandemic (Bille *et al.*, 2005; Derbise *et al.*, 2007; Gonzalez *et al.*, 2002).

4.6.3.4. Survival of Bacterial Pathogens in their Hosts

Other lysogenic conversion genes appear to allow the lysogen to survive better in the host than non-lysogens. Two examples of this are the *bor* and *sodC1* genes found on lambdoid phages. The phage λ Bor protein is expressed from the prophage, and it gives the lysogen a resistance to being killed by guinea pig serum. Many lambdoid phages carry *bor* homologues, and a plasmid-borne homologue of *bor*, called *iss*, has been implicated as a virulence factor for avian colibacillosis (Johnson *et al.*, 2002, 2006).

The *sodC1* gene of *Salmonella* phage Gifsy-2 encodes a periplasmic superoxide dismutase (Figueroa-Bossi and Bossi, 1999). This enzyme helps protect the bacteria from oxy-radical damage during the oxidative burst that occurs during attack by host macrophage cells (Battistoni, 2003). The *sodC1* gene is thought to be expressed from the uninduced prophage, where it is controlled by the host's PhoPQ two-component system, a system that regulates expression of a number of macrophage survival genes (Golubeva and Slauch, 2006).

4.6.3.5. Type III Secretion Effector Proteins

Many bacterial plant and animal pathogens have type III secretion systems that translocate effector proteins, which typically alter the target cells in some way, from the bacterium into the eukaryotic target cell (Galan and Wolf-Watz, 2006). It has recently been found that such effector proteins are often encoded on temperate bacteriophage genomes. The first of these to be discovered was the SopE2 protein of *Salmonella enterica* (Hardt *et al.*, 1998), whose gene lies in the genome of a P2-like phage that integrates into the *ssrA* (tmRNA) gene (Mirold *et al.*, 1999; Pelludat *et al.*, 2003). The role of SopE2 has been studied in some detail, and it is a GTP exchange factor that activates Cdc42 and Rac1 GTPases. Although the picture is not yet complete, the result of such activation, probably in cooperation with other effector proteins, is modulation of the target cell actin cytoskeleton, disruption of tight cell junctions, increased cell membrane ruffling, and eventually invasion of the target cell by the bacterium (Boyle *et al.*, 2006; Deleu *et al.*, 2006; Patel and Galan, 2006; Williams *et al.*, 2004, and references therein).

Since the discovery of SopE, numerous other phage-encoded effector proteins have been found. These include the following effector proteins that contribute to *Salmonella* and *E. coli* virulence in as yet poorly understood ways: *S. enterica* GogB and GtgE, which are encoded by Gifsy-1 and Gifsy-2 lambdoid prophages, respectively (Coombes *et al.*, 2005; Ho *et al.*, 2002), and EspI, EspJ, and EspK of *E. coli* O157:H7 prophages BP-4795, CP-933-U, and CP-933P, respectively (Creuzburg *et al.*, 2005; Dahan *et al.*, 2005; Vlisidou *et al.*, 2006). In addition, putative virulence-related effector proteins appear to be encoded by prophages in the insect pathogens *Serratia entomophila* and *Photorhabdus asymbiotica* (Hurst *et al.*, 2004; Yang *et al.*, 2006). Finally, bioinformatic analysis has identified putative effector loci in about half of *E. coli* O157:H7 Sakai prophages (Tobe *et al.*, 2006). These are likely to be only the tip of the iceberg, and many more phage-encoded effector proteins will likely be discovered.

4.6.3.6. Diseases caused by *Streptococcus pyogenes*

Bacterial virulence genes encoded on prophages are by no means limited to the lambdoid and P2-like phages of enteric bacterial pathogens. For example, the Gram-positive pathogen *S. pyogenes* is the cause of many different severe human diseases, including scarlet fever, toxic shock syndrome, puerperal sepsis, impetigo, necrotizing fasciitis, and post-infection sequelae of acute rheumatic fever. Bacteria in this species are known to harbor as many as six Caudovirales prophages, and these prophages carry a number of the

important *S. pyogenes* virulence factors (Banks *et al.*, 2002; Beres *et al.*, 2006). These temperate phage-encoded virulence factors have been implicated in a number of the features of *S. pyogenes*-caused disease. In particular, prophage genes have been implicated in the disease specificity of puerperal sepsis (Green *et al.*, 2005), in several outbreaks of acute rheumatic fever (Smoot *et al.*, 2002), and in the recent emergence of severely infectious strains since about 1980 (Beres *et al.*, 2004, 2006; Aziz *et al.*, 2005; Sitkiewicz *et al.*, 2006). Some of these virulence factors have been characterized as secreted phospholipase A2 and nucleases, as well as pyrogenic exotoxin superantigens. The lateral exchange of these virulence genes, mediated by phage infection, has clearly been a very important factor in the diversification of *S. pyogenes*.

4.6.4. Other Kinds of 'bacterial fitness modules'?

At present the ways by which prophages convert their hosts, as described earlier, seem skewed toward mechanisms that give pathogens some advantage in their hosts. It seems likely to us that this is a biased view, colored by the fact that human pathogens are under much more intense study than other bacteria, and that temperate phage-borne genes or gene modules will be found that confer a fitness benefit to the bacterial host in virtually any ecological niche. Indeed, prophages are known that affect sporulation of *Clostridium* (Stewart and Johnson, 1977) and cell-cell aggregation of *Streptococcus thermophilus* (Neve *et al.*, 2003), but the mechanisms of these phenomena have not been studied in detail. APSE-2-like (lambdoid) phages are consistently present within the secondary bacterial endosymbiont *Hamiltonia defensa*, which in turn is harbored in bacteriocyte cells of the pea aphid. The *ctdB* gene, located in the same position of these phages' genomes as Shiga toxin genes in *stx*-converting phages, encodes cytolethal distending toxin, and it has been proposed that this toxin might be responsible in part for the insect's defense against eukaryotic parasites (Moran *et al.*, 2005). Molecular mechanisms here have not yet been studied, but this could be a case of a phage increasing the fitness of a bacterium that in turn increases the fitness of the insect.

4.6.5. Gene Regulatory Interactions Between Prophage and Host

It is interesting to note that there are a number of examples of phage moron and lysogenic conversion gene expression that are controlled by host regulatory proteins. These include control of *E. coli* phage 186 *tum* gene,

Mu *mom* gene, and H-19B *stx* genes by the LexA, OxyR, and Fur regulators (Brumby *et al.*, 1996; Calderwood and Mekalanos, 1987; Hattman and Sun, 1997); the *Corynebacterium diphtheriae* prophage β toxin gene by the DtxR regulator (Oram *et al.*, 2002); the *Klebsiella oxytoca* phage ϕKO2 genes *22* and *23* genes by the host LexA repressor (Casjens *et al.*, 2004a); and the *Salmonella* phage Gifsy-2 *sodC1* gene by the host PhoPQ two component regulatory system (Golubeva and Slauch, 2006). These, along with the repression of host genes by phage λ repressor (discussed earlier), suggest an evolutionary interaction between the host and phage where the host regulators and phage regulatory targets co-evolve to optimally supply the lysogenic conversion protein to its host.

4.6.7. Why do Prophages Bother with Lysogenic Conversion?

Potentially lethal prophages that carry genes useful to the host bacteria present the bacterium with a conundrum – are they good or bad? This must have important long-range evolutionary implications, but these are necessarily complex and remain poorly understood. In general, it appears that the prophages often (always?) bring genes with them that give the bacterial host some (short-term?) advantage over non-lysogens. This could keep the host from developing more efficient mechanisms to remove prophages or prevent prophage accumulation. Their location on phage genomes also allows such genes to rapidly spread through the bacterial population, so this could possibly be seen as a bacterial mechanism for spreading useful genes or maintaining species diversity; however, prophages are 'time bombs' in bacterial genomes, so it seems more likely that phages do this for selfish reasons. Since the prophage is replicated with the bacterial chromosome, it is to the phage's advantage to give its host an advantage. From the perspective of the bacterium it should, in the long term, be advantageous to rid itself of the prophage but perhaps retain the useful prophage genes. This prospect is discussed in more detail in the following section.

4.6.7.1. Moron Deposition into the Host Genome?

Although a number of phage lysogenic conversion genes have been identified and their functions studied, much less is known about where these genes originate, whether they are of long-term or only short-term evolutionary advantage to the phages that carry them, and whether or not over the long run they can be appropriated and become a permanent part of the genome of the bacterial host.

The last of these questions is the easiest to approach. Can we identify prophage genes that appear to be in the process of being appropriated by the host? Many of the individual genes in defective prophages that are no longer functional phage genomes and are in the process of evolutionary decay remain completely functional and so could be in the process of incorporation into the host bacterial genome (reviewed in part by Casjens, 2003). But are they useful to the host, or will they continue to decay with the rest of the prophage and eventually disappear? The Shiga toxin of *Shigella dysenteriae* constitutes one of the better cases for prophage genes being caught in mid-assimilation. These toxin genes are arranged in exactly the same way as homologous genes on *stx*-converting phages and lie in similar juxtaposition to lambdoid phage-like genes as in those phages, strongly suggesting that they are derived from a highly deleted lambdoid prophage (McDonough and Butterton, 1999; Unkmeir and Schmidt, 2000). This is shown in Figure 4.2 (see also color plate after p. 174), along with the fact that this prophage relic is not present in the syntenic regions of other *Shigella* species or their close *E. coli* relatives. This defective prophage retains the integrase and excisionase genes, a lambdoid-like homologous recombination gene region, and lysis genes in addition to the toxin gene region, and contains several insertion sequences. The toxin and lysis proteins are all between 90% and 100% identical to homologous genes on known stx-converting phages. Examination of a large number of *S. dysenteriae* 1 strains isolated on different continents and at widely different times in the 20th century all contain nearly identical defective prophages at this location (Greco *et al.*, 2004). Thus, since the Shiga toxin genes have an important role in the *S. dysenteriae* disease process, and since all *S. dysenteriae* 1 cells are descendents of a common ancestor with the same stx-converting prophage at this location, it is eminently reasonable to suppose that the toxin genes (and the lysis genes that might be required for toxin release from cells) are evolutionarily advantageous and are under selection to be retained as the remainder of the prophage disappears.

There are numerous cases where similar arguments for moron assimilation can be made. Among these are the phage-like gene transfer agents and tail-like bacteriocins that seem to be phage virion structural proteins appropriated for the bacterium's purposes (reviewed by Casjens, 2003; Hendrix and Casjens, 2006; Lang and Beatty, 2007; Nakayama *et al.*, 2000). In addition, homologues to the *Salmonella* effector proteins SopE (discussed earlier) and SspH2 (reviewed by Brüssow *et al.*, 2004) are encoded by what appear to be highly degraded prophages, although their current gene content and order do not match known extant temperate phages as neatly as does the *S. dysenteriae* toxin gene case.

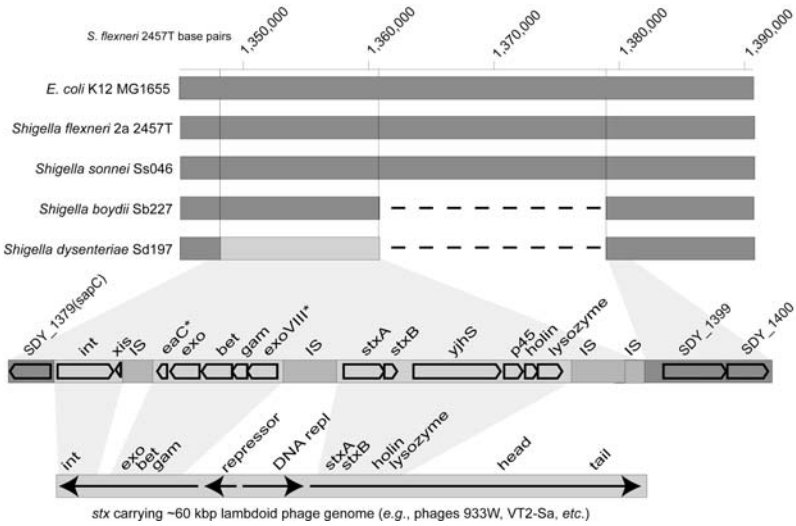


Figure 4.2. The *S. dysenteriae* strain Sd197 Shiga toxin genes lie within a highly deleted defective lambdoid prophage. The red bars at top indicate syntenic regions between the genomes of *E. coli* K12 and the four *Shigella* species (Deng *et al.*, 2002; Yang *et al.*, 2005); these ignore insertion sequence (IS) and other small differences. The dashed black lines indicate DNA that is not present here in *S. boydii* and *S. dysenteriae*. In addition, about 13 kb of DNA present in the other isolates is missing and replaced by putative prophage DNA (light green) in *S. dysenteriae* (McDonough and Butterton, 1999). The prophage replaces the possibly important genes *sapA* and *sapB* (part of a six-gene oligopeptide ABC type permease operon) and glutamine synthetase gene; these are found as SDY_1638, SDY_1639, and SDY_1947, respectively, elsewhere in the *dysenteriae* genome. In the middle, this prophage region is expanded, the predicted genes are indicated as open arrows and the insertion sequences (IS) are denoted in blue. Immediately above the expanded region gene names are shown; all have 'SDY' names, but homologous phage gene names are shown where appropriate. At bottom, a typical *stx*-converting phage genome is shown where black arrows indicate the major phage operons. Regions of homology between the three sections are indicated by light green trapezoids.

Where do phages acquire lysogenic conversion genes? This is a much more difficult question. A number of cases have been noticed in which temperate phage moron and/or lysogenic conversion genes have homologues in non-phage contexts of bacterial genomes. For example, the *E. coli* phage P2 Z/*fun* lysogenic conversion gene, which is responsible for exclusion of phage T5, has an approximately 60% identical homologue in an apparently non-prophage location in *Neisseria meningitidis* chromosome (Nilsson *et al.*, 2004). Unfortunately, it is not possible at present to know whether such

genes are the source from which the phage gene was obtained, or whether they are phage genes that have been fully assimilated.

4.7. SUMMARY AND PROSPECTS

Bacteriophages are the most abundant and diverse organisms on Earth, and they clearly have many different, very important, and long-term effects on the evolution of their bacterial hosts. When bacteria evolve to occupy a new pathogenic niche, it very often involves the horizontal acquisition of virulence genes, and there is no doubt that phages are major players in such transfer. As we have discussed, phages themselves have evolutionary mechanisms to exchange genetic material horizontally, and phages and bacteria also exchange genetic material. In theory, the gene pools of all tailed phages are in contact, and these phage genomes in turn are in contact with bacterial genomes. Thus, phages are a major mediator of horizontal exchange among bacteria by transduction and by moron carriage and so are important facilitators of pathogen generation and evolution. In addition to mediating genetic transfer, many temperate phages have evolved to play a more direct role in bacterial pathogenesis in which they carry and express genes that augment the virulence and aid in the survival of their bacterial hosts when the phage genome is integrated into the bacterial chromosome and not growing lytically. Many critical virulence factors and toxins are made in a controlled fashion from genes that lie in prophages. The detailed evolutionary reasons for this often-repeated relationship are certainly complex and not fully understood, but it appears to be advantageous for both the phage and its host since it has been re-invented so many times. We are currently in the midst of a rush of discoveries of such relationships, and there will be many more such findings over the next few years. More difficult, but very important, will be attaining an understanding why this relationship is so common. Is the ability of prophages to spread horizontally through a population important in the evolution of bacterial pathogens, or is it a purely selfish property of the phages themselves? Where do virulence morons come from? Why do bacteria allow themselves to be dependent upon prophage genes? Questions like these will be excellent grist for the pathogenesis research mill for years to come.

REFERENCES

Allison, G. E., and Verma, N. K. (2000). Serotype-converting bacteriophages and O-antigen modification in *Shigella flexneri*. *Trends Microbiol*, **8**, 17–23.

Altschul, S. F., Madden, T. L., Schaffer, A. A., *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389–402.

Angly, F. E., Felts, B., Breitbart, M., *et al.* (2006). The marine viromes of four oceanic regions. *PLoS Biol*, **4**, e368.

Aziz, R. K., Edwards, R. A., Taylor, W. W., *et al.* (2005). Mosaic prophages with horizontally acquired genes account for the emergence and diversification of the globally disseminated M1T1 clone of *Streptococcus pyogenes*. *J Bacteriol*, **187**, 3311–8.

Baker, M. L., Jiang, W., Rixon, F. J., and Chiu, W. (2005). Common ancestry of herpesviruses and tailed DNA bacteriophages. *J Virol*, **79**, 14967–70.

Banks, D. J., Beres, S. B., and Musser, J. M. (2002). The fundamental contribution of phages to GAS evolution, genome diversification and strain emergence. *Trends Microbiol*, **10**, 515–21.

Battistoni, A. (2003). Role of prokaryotic Cu, Zn superoxide dismutase in pathogenesis. *Biochem Soc Trans*, **31**, 1326–9.

Beres, S. B., Sylva, G. L., Sturdevant, D. E., *et al.* (2004). Genome-wide molecular dissection of serotype M3 group A *Streptococcus* strains causing two epidemics of invasive infections. *Proc Natl Acad Sci USA*, **101**, 11833–8.

Beres, S. B., Richter, E. W., Nagiec, M. J., *et al.* (2006). Molecular genetic anatomy of inter- and intraserotype variation in the human bacterial pathogen group A *Streptococcus*. *Proc Natl Acad Sci USA*, **103**, 7059–64.

Bergh, O., Borsheim, K. Y., Bratbak, G., and Heldal, M. (1989). High abundance of viruses found in aquatic environments. *Nature*, **340**, 467–8.

Bille, E., Zahar, J. R., Perrin, A., *et al.* (2005). A chromosomally integrated bacteriophage in invasive meningococci. *J Exp Med*, **201**, 1905–13.

Bose, M., and Barber, R. D. (2006). Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. *In Silico Biol*, **6**, 223–7.

Botstein, D. (1980). A theory of modular evolution in bacteriophages. *Ann N Y Acad Sci*, **354**, 484–91.

Boyd, E. F., and Brüssow, H. (2002). Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. *Trends Microbiol*, **10**, 521–9.

Boyle, E. C., Brown, N. F., and Finlay, B. B. (2006). *Salmonella enterica* serovar Typhimurium effectors SopB, SopE, SopE2 and SipA disrupt tight junction structure and function. *Cell Microbiol*, **8**, 1946–57.

Breitbart, M., Felts, B., Kelley, S., *et al.* (2004a). Diversity and population structure of a near-shore marine-sediment viral community. *Proc Biol Sci*, **271**, 565–74.

Breitbart, M., Miyake, J. H., and Rohwer, F. (2004b). Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol Lett*, **236**, 249–56.

Brumby, A. M., Lamont, I., Dodd, I. B., and Egan, J. B. (1996). Defining the SOS operon of coliphage 186. *Virology*, **219**, 105–14.

Brüssow, H., Canchaya, C., and Hardt, W. D. (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol Rev*, **68**, 560–602.

Calderwood, S. B., and Mekalanos, J. J. (1987). Iron regulation of Shiga-like toxin expression in *Escherichia coli* is mediated by the fur locus. *J Bacteriol*, **169**, 4759–64.

Campbell, A. M. (2002). Preferential orientation of natural lambdoid prophages and bacterial chromosome organization. *Theor Popul Biol*, **61**, 503–7.

Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M. L., and Brüssow, H. (2003a). Phage as agents of lateral gene transfer. *Curr Opin Microbiol*, **6**, 417–24.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., and Brüssow, H. (2003b). Prophage genomics. *Microbiol Mol Biol Rev*, **67**, 238–76.

Canchaya, C., Fournous, G., and Brüssow, H. (2004). The impact of prophages on bacterial chromosomes. *Mol Microbiol*, **53**, 9–18.

Casjens, S. (2003). Prophages in bacterial genomics: What have we learned so far? *Mol Microbiol*, **249**, 277–300.

Casjens, S. R. (2005). Comparative genomics and evolution of the tailed-bacteriophages. *Curr Opin Microbiol*, **8**, 451–8.

Casjens, S., and Hendrix, R. (2005). Bacteriophages and the bacterial genome. In Higgins, N. P. (Ed.). The bacterial chromosome. Washington, DC: ASM Press.

Casjens, S., Hatfull, G., and Hendrix, R. (1992). Evolution of dsDNA tailed-bacteriophage genomes. *Sem Virol*, **3**, 383–97.

Casjens, S., Gilcrease, E. B., Huang, W. M., *et al.* (2004a). The pKO2 linear plasmid prophage of *Klebsiella oxytoca*. *J Bacteriol*, **186**, 1818–32.

Casjens, S., Winn-Stapley, D., Gilcrease, E., *et al.* (2004b). The chromosome of *Shigella flexneri* bacteriophage Sf6: complete nucleotide sequence, genetic mosaicism, and DNA packaging. *J Mol Biol*, **339**, 379–94.

Casjens, S. R., Gilcrease, E. B., Winn-Stapley, D. A., *et al.* (2005). The generalized transducing *Salmonella* bacteriophage ES18: complete genome sequence and DNA packaging strategy. *J Bacteriol*, **187**, 1091–104.

Chen, Y., Golding, I., Sawai, S., Guo, L., and Cox, E. C. (2005). Population fitness and the regulation of *Escherichia coli* genes by bacterial viruses. *PLoS Biol*, **3**, e229.

Coombes, B. K., Wickham, M. E., Brown, N. F., *et al.* (2005). Genetic and molecular analysis of GogB, a phage-encoded type III-secreted substrate

in *Salmonella enterica* serovar Typhimurium with autonomous expression from its associated phage. *J Mol Biol*, **348**, 817–30.

Creuzburg, K., Köhler, B., Hempel, H., *et al.* (2005). Genetic structure and chromosomal integration site of the cryptic prophage CP-1639 encoding Shiga toxin 1. *Microbiology*, **151**, 941–50.

Dahan, S., Wiles, S., La Ragione, R. M., *et al.* (2005). EspJ is a prophage-carried type III effector protein of attaching and effacing pathogens that modulates infection dynamics. *Infect Immun*, **73**, 679–86.

Danovaro, R., and Serresi, M. (2000). Viral density and virus-to-bacterium ratio in deep-sea sediments of the Eastern Mediterranean. *Appl Environ Microbiol*, **66**, 1857–61.

De Haan, L., and Hirst, T. R. (2004). Cholera toxin: a paradigm for multifunctional engagement of cellular mechanisms. *Mol Membr Biol*, **21**, 77–92.

Deleu, S., Choi, K., Reece, J. M., and Shears, S. B. (2006). Pathogenicity of *Salmonella*: SopE-mediated membrane ruffling is independent of inositol phosphate signals. *FEBS Lett*, **580**, 1709–15.

Deng, W., Burland, V., Plunkett, G. 3rd, *et al.* (2002). Genome sequence of *Yersinia pestis* KIM. *J Bacteriol*, **184**, 4601–11.

Derbise, A., Chenal-Francisque, V., Pouillot, F., *et al.* (2007). A horizontally acquired filamentous phage contributes to the pathogenicity of the plague bacillus. *Mol Microbiol*, **63**, 1145–57.

Ebel-Tsipis, J., Botstein, D., and Fox, M. S. (1972). Generalized transduction by phage P22 in *Salmonella typhimurium*. I. Molecular origin of transducing DNA. *J Mol Biol*, **71**, 433–48.

Edlin, G., Lin, L., and Kudrna, R. (1975). Lambda lysogens of *E. coli* reproduce more rapidly than non-lysogens. *Nature*, **255**, 735–7.

Edwards, R. A., and Rohwer, F. (2005). Viral metagenomics. *Nat Rev Microbiol*, **3**, 504–10.

Fauquet, C. M., Mayo, M. A., Maniloff, J., Desselberger, U., and Ball, L. A. (Eds.) (2005). *Virus taxonomy: eighth report of the International Committee on Taxonomy of Viruses*. Amsterdam, The Netherlands, Elsevier Academic Press.

Figueroa-Bossi, N., and Bossi, L. (1999). Inducible prophages contribute to *Salmonella* virulence in mice. *Mol Microbiol*, **33**, 167–76.

Fokine, A., Leiman, P. G., Shneider, M. M., *et al.* (2005). Structural and functional similarities between the capsid proteins of bacteriophages T4 and HK97 point to a common ancestry. *Proc Natl Acad Sci USA*, **102**, 7163–8.

Forterre, P. (2006). Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: a hypothesis for the origin of cellular domain. *Proc Natl Acad Sci USA*, **103**, 3669–74.

Fouts, D. E. (2006). Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res*, **34**, 5839–51.

Freeman, V. (1951). Studies on the virulence of bacteriophage-infected strains of *Corynebacterium diphtheriae*. *J Bacteriol*, **61**, 675–88.

Fukazawa, Y., and Hartman, P. (1964). A P22 bacteriophage mutant defective in antigenic conversion. *Virology*, **23**, 279–83.

Galan, J. E., and Wolf-Watz, H. (2006). Protein delivery into eukaryotic cells by type III secretion machines. *Nature*, **444**, 567–73.

Garza, D. R., and Suttle, C. A. (1998). The effect of Cyanophages on the mortality of *Synechococcus* spp., and selection for UV resistant viral communities. *Microb Ecol*, **36**, 281–92.

Georgopoulos, C. P., Hendrix, R. W., Casjens, S. R., and Kaiser, A. D. (1973). Host participation in bacteriophage lambda head assembly. *J Mol Biol*, **76**, 45–60.

Gilcrease, E. B., Winn-Stapley, D. A., Hewitt, F. C., Joss, L., and Casjens, S. R. (2005). Nucleotide sequence of the head assembly gene cluster of bacteriophage L and decoration protein characterization. *J Bacteriol*, **187**, 2050–7.

Golubeva, Y. A., and Slauch, J. M. (2006). *Salmonella enterica* serovar Typhimurium periplasmic superoxide dismutase SodCI is a member of the PhoPQ regulon and is induced in macrophages. *J Bacteriol*, **188**, 7853–61.

Gonzalez, M. D., Lichtensteiger, C. A., Caughlan, R., and Vimr, E. R. (2002). Conserved filamentous prophage in *Escherichia coli* O18:K1:H7 and *Yersinia pestis* biovar Orientalis. *J Bacteriol*, **184**, 6050–5.

Greco, K. M., McDonough, M. A., and Butterton, J. R. (2004). Variation in the Shiga toxin region of 20th-century epidemic and endemic *Shigella dysenteriae* 1 strains. *J Infect Dis*, **190**, 330–4.

Green, N. M., Zhang, S., Porcella, S. F., *et al*. (2005). Genome sequence of a serotype M28 strain of group a *Streptococcus*: potential new insights into puerperal sepsis and bacterial disease specificity. *J Infect Dis*, **192**, 760–70.

Hambly, E., and Suttle, C. A. (2005). The viriosphere, diversity, and genetic exchange within phage communities. *Curr Opin Microbiol*, **8**, 444–50.

Hardt, W. D., Urlaub, H., and Galan, J. E. (1998). A substrate of the centisome 63 type III protein secretion system of *Salmonella typhimurium* is encoded by a cryptic bacteriophage. *Proc Natl Acad Sci USA*, **95**, 2574–9.

Hatfull, G. F., Pedulla, M. L., Jacobs-Sera, D., *et al*. (2006). Exploring the mycobacteriophage metaproteome: phage genomics as an educational platform. *PLoS Genet*, **2**, e92.

Hattman, S., and Sun, W. (1997). *Escherichia coli* OxyR modulation of bacteriophage Mu *mom* expression in *dam*$^+$ cells can be attributed to its ability to bind hemimethylated P_{mom} promoter DNA. *Nucleic Acids Res*, **25**, 4385–8.

Hendrix, R., and Casjens, S. (2006). Bacteriophage λ and its genetic neighborhood. In Calendar, R. (Ed.) The bacteriophages, (2nd. ed.). New York: Oxford.

Hendrix, R. W. (2002). Bacteriophages: evolution of the majority. *Theor Popul Biol*, **61**, 471–80.

Hendrix, R. W. (2003). Bacteriophage genomics. *Curr Opin Microbiol*, **6**, 506–11.

Hendrix, R. W., Lawrence, J. G., Hatfull, G. F., and Casjens, S. (2000). The origins and ongoing evolution of viruses. *Trends Microbiol*, **8**, 504–8.

Herold, S., Karch, H., and Schmidt, H. (2004). Shiga toxin-encoding bacteriophages – genomes in motion. *Int J Med Microbiol*, **294**, 115–21.

Hershey, A. D., and Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol*, **36**, 39–56.

Hershey, A. D., and Rotman, R. (1949). Genetic recombination between host range and plaque-type mutants of bacteriophage in single bacterial cells. *Genetics*, **34**, 44–55.

Hewson, I., O'Neil, I., Fuhrman, J. A., and Dennison, W. C. (2001). Virus-like particle distribution and abundance in sediments and overlying waters along eutrophication gradients in two subtropical estuaries. *Limnol Oceanogr*, **46**, 1734–46.

Ho, T. D., Figueroa-Bossi, N., Wang, M., *et al.* (2002). Identification of GtgE, a novel virulence factor encoded on the Gifsy-2 bacteriophage of *Salmonella enterica* serovar Typhimurium. *J Bacteriol*, **184**, 5234–9.

Hurst, M. R., Glare, T. R., and Jackson, T. A. (2004). Cloning *Serratia entomophila* antifeeding genes – a putative defective prophage active against the grass grub *Costelytra zealandica*. *J Bacteriol*, **186**, 5116–28.

Iguchi, A., Iyoda, S., Terajima, J., Watanabe, H., and Osawa, R. (2006). Spontaneous recombination between homologous prophage regions causes large-scale inversions within the *Escherichia coli* O157:H7 chromosome. *Gene*, **372**, 199–207.

Jiang, S. C., and Paul, J. H. (1998). Gene transfer by transduction in the marine environment. *Appl Environ Microbiol*, **64**, 2780–7.

Jiang, W., Li, Z., Zhang, Z., *et al.* (2003). Coat protein fold and maturation transition of bacteriophage P22 seen at subnanometer resolutions. *Nat Struct Biol*, **10**, 131–5.

Johnson, T., Giddings, C., Horne, S., *et al.* (2002). Location of increased serum survival gene and selected virulence traits on a conjugative R plasmid in an avian *Escherichia coli* isolate. *Avian Dis*, **46**, 342–542.

Johnson, T. J., Siek, K. E., Johnson, S. J., and Nolan, L. K. (2006). DNA sequence of a ColV plasmid and prevalence of selected plasmid-encoded virulence genes among avian *Escherichia coli* strains. *J Bacteriol*, **188**, 745–58.

Juhala, R. J., Ford, M. E., Duda, R. L., *et al*. (2000). Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. *J Mol Biol*, **299**, 27–51.

Killmann, H., Braun, M., Herrmann, C., and Braun, V. (2001). FhuA barrel-cork hybrids are active transporters and receptors. *J Bacteriol*, **183**, 3476–87.

Koonin, E. V., Senkevich, T. G., and Dolja, V. V. (2006). The ancient Virus World and evolution of cells. *Biol Direct*, **1**, 29.

Korres, H., and Verma, N. K. (2006). Identification of essential loops and residues of glucosyltransferase V (GtrV) of *Shigella flexneri*. *Mol Membr Biol*, **23**, 407–19.

Lang, A. S., and Beatty, J. T. (2007). Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol*, **15**, 54–62.

Lawrence, J. G., Hendrix, R. W., and Casjens, S. (2001). Where are the bacterial pseudogenes? *Trends Microbiol*, **9**, 535–40.

Lindell, D., Sullivan, M. B., Johnson, Z. I., *et al*. (2004). Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc Natl Acad Sci USA*, **101**, 11013–8.

Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M., and Chisholm, S. W. (2005). Photosynthesis genes in marine viruses yield proteins during host infection. *Nature*, **438**, 86–9.

Luria, S. E., and Delbrück, M. (1943). Mutation of bacteria from virus sensitivity to virus resistance. *Genetics*, **28**, 491–511.

Mann, N. H., Cook, A., Millard, A., Bailey, S., and Clokie, M. (2003). Marine ecosystems: bacterial photosynthesis genes in a virus. *Nature*, **424**, 741.

Markine-Goriaynoff, N., Gillet, L., Van Etten, J. L., *et al*. (2004). Glycosyltransferases encoded by viruses. *J Gen Virol*, **85**, 2741–54.

Matsumoto, M., Ichikawa, N., Tanaka, S., Morita, T., and Matsushiro, A. (1985). Molecular cloning of the ϕ80 adsorption-inhibiting *cor* gene. *Jpn J Genet*, **60**, 475–83.

McDonough, M. A., and Butterton, J. R. (1999). Spontaneous tandem amplification and deletion of the Shiga toxin operon in *Shigella dysenteriae* 1. *Mol Microbiol*, **34**, 1058–69.

Mehta, P., Casjens, S., and Krishnaswamy, S. (2004). Analysis of the lambdoid prophage element e14 in the *E. coli* K-12 genome. *BMC Microbiol*, **4**, 4.

Millard, A., Clokie, M. R., Shub, D. A., and Mann, N. H. (2004). Genetic organization of the *psbAD* region in phages infecting marine *Synechococcus* strains. *Proc Natl Acad Sci USA*, **101**, 11007–12.

Mirold, S., Rabsch, W., Rohde, M., *et al.* (1999). Isolation of a temperate bacteriophage encoding the type III effector protein SopE from an epidemic *Salmonella typhimurium* strain. *Proc Natl Acad Sci USA*, **96**, 9845–50.

Morais, M. C., Choi, K. H., Koti, J. S., *et al.* (2005). Conservation of the capsid structure in tailed dsDNA bacteriophages: the pseudoatomic structure of phi29. *Mol Cell*, **18**, 149–59.

Moran, N. A., Degnan, P. H., Santos, S. R., Dunbar, H. E., and Ochman, H. (2005). The players in a mutualistic symbiosis: insects, bacteria, viruses, and virulence genes. *Proc Natl Acad Sci USA*, **102**, 16919–26.

Morgan, G., Hatfull, G., Casjens, S., and Hendrix, R. (2002). Bacteriophage Mu genome sequence: analysis and comparison with Mu-like prophages in *Haemophilus, Neisseria* and *Deinococcus*. *J Mol Biol*, **317**, 337–59.

Nakagawa, I., Kurokawa, K., Yamashita, A., *et al.* (2003). Genome sequence of an M3 strain of *Streptococcus pyogenes* reveals a large-scale genomic rearrangement in invasive strains and new insights into phage evolution. *Genome Res*, **13**, 1042–55.

Nakayama, K., Takashima, K., Ishihara, H., *et al.* (2000). The R-type pyocin of *Pseudomonas aeruginosa* is related to P2 phage, and the F-type is related to lambda phage. *Mol Microbiol*, **38**, 213–31.

Neve, H., Freudenberg, W., Diestel-Feddersen, F., Ehlert, R., and Heller, K. J. (2003). Biology of the temperate *Streptococcus thermophilus* bacteriophage TP-J34 and physical characterization of the phage genome. *Virology*, **315**, 184–94.

Newton, G. J., Daniels, C., Burrows, L. L., *et al.* (2001). Three-component-mediated serotype conversion in *Pseudomonas aeruginosa* by bacteriophage D3. *Mol Microbiol*, **39**, 1237–47.

Nilsson, A. S., Karlsson, J. L., and Haggard-Ljungquist, E. (2004). Site-specific recombination links the evolution of P2-like coliphages and pathogenic enterobacteria. *Mol Biol Evol*, **21**, 1–13.

Ochman, H., Lerat, E., and Daubin, V. (2005). Examining bacterial species under the specter of gene transfer and exchange. *Proc Natl Acad Sci USA*, **102**(Suppl 1), 6595–9.

Oram, D. M., Avdalovic, A., and Holmes, R. K. (2002). Construction and characterization of transposon insertion mutations in *Corynebacterium diphtheriae* that affect expression of the diphtheria toxin repressor (DtxR). *J Bacteriol*, **184**, 5723–32.

Patel, J. C., and Galan, J. E. (2006). Differential activation and function of Rho GTPases during *Salmonella*–host cell interactions. *J Cell Biol*, **175**, 453–63.

Pedulla, M. L., Ford, M. E., Houtz, J. M., *et al.* (2003). Origins of highly mosaic mycobacteriophage genomes. *Cell*, **113**, 171–82.

Pelludat, C., Mirold, S., and Hardt, W. D. (2003). The SopEφ phage integrates into the *ssrA* gene of *Salmonella enterica* serovar Typhimurium A36 and is closely related to the Fels-2 prophage. *J Bacteriol*, **185**, 5182–91.

Penner, M., Morad, I., Snyder, L., and Kaufmann, G. (1995). Phage T4-coded Stp: double-edged effector of coupled DNA and tRNA-restriction systems. *J Mol Biol*, **249**, 857–68.

Piuri, M., and Hatfull, G. F. (2006). A peptidoglycan hydrolase motif within the mycobacteriophage TM4 tape measure protein promotes efficient infection of stationary phase cells. *Mol Microbiol*, **62**, 1569–85.

Quinones, M., Davis, B. M., and Waldor, M. K. (2006a). Activation of the *Vibrio cholerae* SOS response is not required for intestinal cholera toxin production or colonization. *Infect Immun*, **74**, 927–30.

Quinones, M., Kimsey, H. H., Ross, W., Gourse, R. L., and Waldor, M. K. (2006b). LexA represses CTXφ transcription by blocking access of the alpha C-terminal domain of RNA polymerase to promoter DNA. *J Biol Chem*, **281**, 39407–12.

Sandkvist, M. (2001). Type II secretion and pathogenesis. *Infect Immun*, **69**, 3523–35.

Silver-Mysliwiec, T. H., and Bramucci, M. G. (1990). Bacteriophage-enhanced sporulation: comparison of spore-converting bacteriophages PMB12 and SP10. *J Bacteriol*, **172**, 1948–53.

Sitkiewicz, I., Nagiec, M. J., Sumby, P., *et al.* (2006). Emergence of a bacterial clone with enhanced virulence by acquisition of a phage encoding a secreted phospholipase A2. *Proc Natl Acad Sci USA*, **103**, 16009–14.

Smoot, J. C., Barbian, K. D., Van Gompel, J. J., *et al.* (2002). Genome sequence and comparative microarray analysis of serotype M18 group A *Streptococcus* strains associated with acute rheumatic fever outbreaks. *Proc Natl Acad Sci USA*, **99**, 4668–73.

Stent, G. (1963). *Molecular biology of bacterial viruses*, San Francisco: W. H. Freeman.

Sternberg, N., and Weisberg, R. (1977). Packaging of coliphage lambda DNA. II. The role of the gene D protein. *J Mol Biol*, **117**, 733–59.

Stewart, A. W., and Johnson, M. G. (1977). Increased numbers of heat-resistant spores produced by two strains of *Clostridium perfringens* bearing temperate phage s9. *J Gen Microbiol*, **103**, 45–50.

Sullivan, M. B., Lindell, D., Lee, J. A., *et al.* (2006). Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol*, **4**, e234.

Summers, W. C. (1999). *Félix d'Herelle and the origins of molecular biology*, New Haven, CT: Yale University Press.

Susskind, M. M., and Botstein, D. (1978). Molecular genetics of bacteriophage P22. *Microbiol Rev*, **42**, 385–413.

Suttle, C. A. (2005). Viruses in the sea. *Nature*, **437**, 356–61.

Tang, L., Gilcrease, E. B., Casjens, S. R., and Johnson, J. E. (2006). Highly discriminatory binding of capsid-cementing proteins in bacteriophage L. *Structure*, **14**, 837–45.

Tobe, T., Beatson, S. A., Taniguchi, H., *et al.* (2006). An extensive repertoire of type III secretion effectors in *Escherichia coli* O157 and the role of lambdoid phages in their dissemination. *Proc Natl Acad Sci USA*, **103**, 14941–6.

Uc-Mass, A., Loeza, E. J., De La Garza, M., *et al.* (2004). An orthologue of the *cor* gene is involved in the exclusion of temperate lambdoid phages. Evidence that Cor inactivates FhuA receptor functions. *Virology*, **329**, 425–33.

Uchida, T., Gill, D. M., and Pappenheimer, A. M., Jr. (1971). Mutation in the structural gene for diphtheria toxin carried by temperate phage. *Nat New Biol*, **233**, 8–11.

Unkmeir, A., and Schmidt, H. (2000). Structural analysis of phage-borne *stx* genes and their flanking sequences in Shiga toxin-producing *Escherichia coli* and *Shigella dysenteriae* type 1 strains. *Infect Immun*, **68**, 4856–64.

Vlisidou, I., Dziva, F., La Ragione, R. M., *et al.* (2006). Role of intimin-tir interactions and the tir-cytoskeleton coupling protein in the colonization of calves and lambs by *Escherichia coli* O157:H7. *Infect Immun*, **74**, 758–64.

Wagner, P. L., and Waldor, M. K. (2002). Bacteriophage control of bacterial virulence. *Infect Immun*, **70**, 3985–93.

Wagner, P. L., Livny, J., Neely, M. N., *et al.* (2002). Bacteriophage control of Shiga toxin 1 production and release by *Escherichia coli*. *Mol Microbiol*, **44**, 957–70.

Waldor, M. K., and Mekalanos, J. J. (1996). Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science*, **272**, 1910–4.

Weigele, P. R., Pope, W. H., Pedulla, M. L., *et al.* (2007). Genomic and structural analysis of Syn9, a cyanophage infecting marine *Prochlorococcus* and *Synechococcus*. *Environ Microbiol*, **9**, 1675–95.

Weinbauer, M. G., Fuks, D., and Peduzzi, P. (1993). Distribution of viruses and dissolved DNA along a coastal trophic gradient in the northern Adriatic sea. *Appl Environ Microbiol*, **59**, 4074–82.

Weitz, J. S., Hartman, H., and Levin, S. A. (2005). Coevolutionary arms races between bacteria and bacteriophage. *Proc Natl Acad Sci USA*, **102**, 9535–40.

West, N. P., Sansonetti, P., Mounier, J., *et al.* (2005). Optimization of virulence functions through glucosylation of *Shigella* LPS. *Science*, **307**, 1313–7.

Whitman, W. B., Coleman, D. C., and Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA*, **95**, 6578–83.

Wikoff, W. R., Liljas, L., Duda, R. L., *et al.* (2000). Topologically linked protein rings in the bacteriophage HK97 capsid. *Science*, **289**, 2129–33.

Williams, C., Galyov, E. E., and Bagby, S. (2004). Solution structure, backbone dynamics, and interaction with Cdc42 of *Salmonella* guanine nucleotide exchange factor SopE2. *Biochemistry*, **43**, 11998–2008.

Williams, K. P. (2002). Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res*, **30**, 866–75.

Yang, F., Yang, J., Zhang, X., *et al.* (2005). Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery. *Nucleic Acids Res*, **33**, 6445–58.

Yang, G., Dowling, A. J., Gerike, U., Ffrench-Constant, R. H., and Waterfield, N. R. (2006). *Photorhabdus* virulence cassettes confer injectable insecticidal activity against the wax moth. *J Bacteriol*, **188**, 2254–61.

Zeidner, G., Bielawski, J. P., Shmoish, M., *et al.* (2005). Potential photosynthesis gene recombination between *Prochlorococcus* and *Synechococcus* via viral intermediates. *Environ Microbiol*, **7**, 1505–13.

CHAPTER 5

Genomic Islands in the Bacterial Chromosome – Paradigms of Evolution in Quantum Leaps

Tobias Ölschläger and Jörg Hacker

5.1. INTRODUCTION

The bacterial genome, that is, the entirety of all genes of a bacterium, was once viewed as a rather stable entity. However, the observation of spreading resistance to antibiotics led to the discovery of extra-chromosomal elements encoding this property. Obviously, plasmids are able to transfer genes from one bacterium to another not only among one species but also from one species to another. Such transfer of genes is not restricted to antibiotic resistance genes. Examples of further traits often encoded by plasmids include resistance to heavy metals and production of toxins.

Another example of mobile genetic elements is phages, the viruses of bacteria. Phages are not just able to infect and finally lyse the bacterial host cell. Certain phages infect and then integrate their whole genome into the bacterial chromosome and thereby become a prophage. This may add another important factor to the property of the infected bacteria. In the case of pathogenic bacteria the production of toxins is frequently encoded by a prophage. A few medically important examples are bacteriophage β of *Corynebacterium diphtheriae* encoding diphtheria toxin, phage C1 of *Clostridium botulinum* coding for the C1 neurotoxin, and phage H-19B of *Escherichia coli*, which harbors the gene for Shiga toxin Stx1 (for a recent review, see Brüssow *et al.*, 2004).

Smaller but still important mobile genetic units are insertion sequence (IS) elements. IS elements mediate DNA rearrangements by transposition, resulting in off/on switching of gene expression by insertion into, and excision from, open reading frames (ORFs), respectively. In addition, IS elements are involved in moving genetic information between the chromosome and plasmids. Two IS elements framing genes form a discrete unit termed a

transposon. Transposons may encode factors mediating resistance to antibiotics or heavy metals or toxin production (e.g., the heat-stable enterotoxin of *E. coli*). They are able to change their location in the chromosome or plasmid and switch between these genetic entities (Bennett, 2004).

Physically linked to mobile DNA elements, such as IS elements, transposons, and conjugative plasmids, are the five classes of mobile integrons. Integrons are genetic units that incorporate exogenous ORFs by site-specific recombination and convert them to functional genes by ensuring their correct expression. They are composed of a gene (*intI*) encoding an integrase belonging to the tyrosine-recombinase, a primary recombination site (*attI*), and an outward-oriented (constitutive) promoter (P_c) that directs transcription of the captured genes (Hall and Collis, 1995). The integron-encoded integrase recombines circularized DNA known as gene cassettes in a RecA-independent manner downstream of P_c at the *attI* site. These gene cassettes generally contain a single gene and an imperfect inverted repeat at the 3′ end of the gene called *attC*. The *attC* sites are nucleotide sequences that function as recognition sites for the site-specific integrase. The gene cassettes in integrons usually encode resistance to antibiotics or antiseptics and are of diverse origins as indicated by the differences in codon usage among gene cassettes in the same integron. Integrons harboring up to eight different resistance cassettes have been characterized. Such mobile integrons not only have been identified in various Gram-negative species but also have been found in Gram-positive bacteria including corynebacteria, aerococci, brevibacteria, and staphylococci (Nandi *et al.*, 2004). However, with the discovery of thousands of gene cassettes associated with integrons in the genomes of environmental species, the importance of these elements clearly extends beyond the phenomenon of antibiotic resistance. The recruitment of integron gene cassettes endows the recipient bacteria with new functions, potentially giving the microorganism an adaptive evolutionary advantage (Mazel, 2006).

5.2. GENOMIC ISLANDS (GEI)

All of these mobile genetic elements are part of the so-called flexible gene pool (Hacker and Carniel, 2001). The other part of the genome is the core genome which encodes for essential functions. Besides the already mentioned parts of the flexible gene pool, there are also ‘genomic islands’ (GEI) (Figure 5.1; Dobrindt *et al.*, 2004). The observation of the spontaneous loss of hemolytic activity by a uropathogenic *E. coli* strain and the

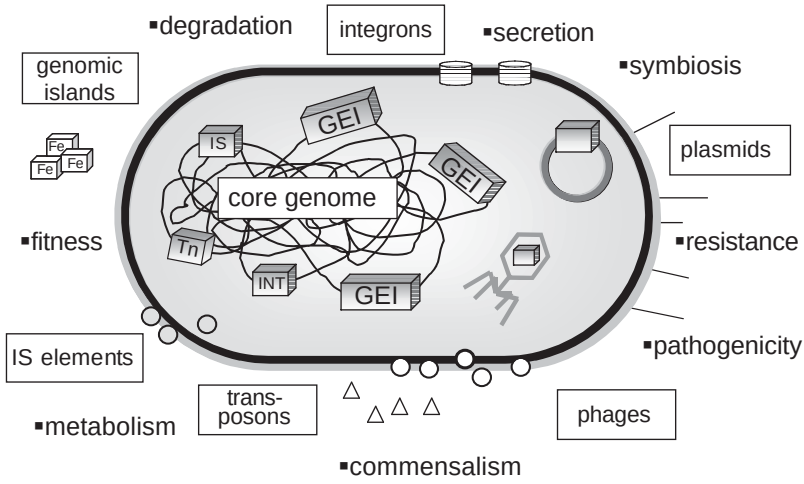


Figure 5.1. Core genome and flexible gene pool of bacteria. The flexible gene pool comprises mobile genetic elements such as genomic islands (GEI), phages, plasmids, integrons (INT), transposons (Tn), and insertion sequence (IS) elements. The flexible gene pool may encode genes for degradation, secretion, symbiosis, resistance, pathogenicity, metabolism, and fitness.

molecular characterization of the genetic event that was responsible for this phenotypic change laid the basis for the concept of 'pathogenicity islands' (PAI). It was demonstrated that the hemolysin-negative mutant had lost not only two determinants encoding α-hemolysin but in addition two large, obviously unstable regions of the bacterial chromosome adjacent to the α-hemolysin gene. These regions were termed PAI and represent a subgroup of GEI (Hacker *et al.*, 1990). In particular, genome sequencing revealed the widespread presence of GEI especially among Gram-negative bacterial species. To date, GEI have been described for more than 30 microbial species.

5.2.1. Characteristics and Composition of GEI

Genomic islands are most often located on the chromosome in both pathogenic and non-pathogenic bacteria. However, they may also be part of plasmids. In contrast to the core genome, GEI exhibit a different G+C content and codon usage. On the chromosome they are frequently inserted into genes for tRNAs (Hochhut *et al.*, 2006b). However, a large-scale polymerase

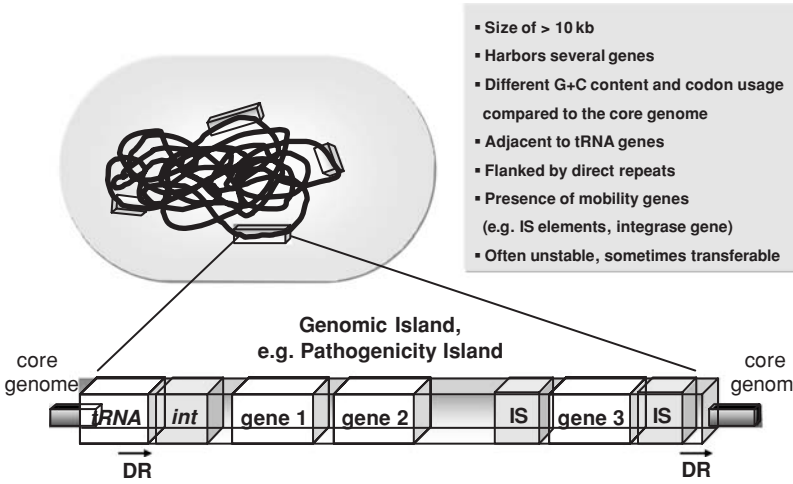


Figure 5.2. Bacterial genomic islands (GEI) are discrete units. Typical features of GEI in Gram-negative bacteria are listed in the box. They encode often an integrase (*int*) and several genes for specific properties. In addition, mobility genes are present as IS elements.

chain reaction (PCR)-based screening revealed that in *E. coli* polycistronic tDNA, highly transcribed tDNA or tDNA encoding tRNA recognizing frequently used codons was generally not targeted by ecto-chromosomal DNA (Germon *et al.*, 2007). The boundaries of GEI are typically formed by direct repeats. GEI usually do not harbor just one gene; rather, they contain whole sets or cassettes of functionally related genes. Therefore, their size exceeds 10 kb. In addition, mobility genes as IS elements and transposons are present on GEI. Finally, an integrase gene is found on many GEI close to the tRNA gene (Figure 5.2; Hochhut *et al.*, 2006b). The latter properties might be responsible for the instability of certain GEI or reflect the acquisition of GEI by horizontal gene transfer. Some GEI harbor genes encoding a restriction-modification system such as the *Yersinia* adhesion pathogenicity island (YAPI) of *Yersinia pseudotuberculosis* and the SCC composite island (SCC-CI) in *Staphylococcus epidermidis* (Collyn *et al.*, 2004; Mongkolrattanothai *et al.*, 2004). The respective restriction-modification systems might increase stability of the encoding GEI. The presence of phage- (e.g., integrase genes) or plasmid-derived genes are indications for the origin of GEI (Hacker and Kaper, 2000). The integration of whole gene blocks by the acquisition of GEI may lead to a dramatic change of properties of the recipient resulting in 'evolution in quantum leaps' (Groisman and Ochman, 1996).

5.2.2. Subgroups of GEI

GEI are not grouped just according to the specific function(s) encoded by a certain GEI but rather according to the lifestyle of the bacterial recipient. In consequence, a certain GEI, such as the yersiniabactin iron uptake-mediating GEI, might be allocated to different GEI groups according to the recipient. The yersiniabactin island might be viewed as a PAI if present in a bacterial pathogen such as *Yersinia* spp., as a saprophytic island (SAI) in commensal *E. coli* or as an ecological island (ECI) if harbored by *Klebsiella* spp. from soil. Examples of symbiosis islands (SYI) are the SYI of *Mesorhizobium loti* encoding the ability of nitrogen fixation and the SYI of *Sinorhizobium fredii* encoding a type III protein secretion system (T3SS). GEI carrying genes mediating resistance toward antibiotics are the most frequent cause in *Staphylococcus* spp. and can be viewed as resistance islands (Table 5.1). These resistance islands are termed SCC*mec* for staphylococcal chromosomal cassette containing the *mecA* gene and are responsible for methicillin resistance by production of the transpeptidase penicillin-binding protein 2′ or 2a, which has decreased affinity for β-lactam antibiotics (Ito *et al.*, 2001). In any case the GEI is maintained in the microbe as long as it provides a selective advantage for the bacterial host. This advantage or increased fitness might be enhanced survival, spread, or transmission resulting from the properties encoded by a GEI. Therefore, such GEI are also termed 'fitness islands' (Hacker and Carniel, 2001).

5.3. PATHOGENICITY ISLANDS (PAI)

5.3.1. Stability and Transfer

The PAI are a subgroup of GEI that is characterized by contributing directly or indirectly to the virulence of a pathogen and to the damage inflicted on the infected host (Hacker and Kaper, 1999). After acquisition, most often PAI seem to have undergone modifications in or deletions of mobility genes resulting in stabilization of the respective PAI. Alteration of the flanking direct repeats is an additional way leading to a more stable association of the PAI with the chromosome. On the other hand, certain PAI are lost at a certain frequency. This is exemplified by PAI I and II of the uropathogenic *E. coli* (UPEC) strain 536, an event that had led to the discovery of PAI. Of the several PAI of strain 536, site-specific excision from the chromosome was demonstrated for four PAI, all flanked by direct repeats. The PAI encoded P4-like integrase genes are required for these deletions (Hochhut *et al.*, 2006b). The frequency of deletion in vitro is influenced by several

Table 5.1. *Examples of subgroups of GEI*

GEI*	Subtype	Localization	Function	Organism	Gram
HPI	PAI	chromosome	iron uptake	*Yersinia* spp.	negative
HPI	SAI	chromosome	iron uptake	commensal *Escherichia coli*	negative
HPI	ECI	chromosome	iron uptake	*Klebsiella* spp.	negative
SCC*mec*	REI	chromosome	methicillin resistance	*Staphylococcus aureus*	positive
SaPI1	PAI	chromosome	toxin production	*Staphylococcus aureus*	positive
PPI-1	PAI	chromosome	iron uptake	*Streptococcus pneumoniae*	positive
pOX1 island	PAI	plasmid	toxin production	*Bacillus anthracis*	positive
mxi/spa island	PAI	plasmid	T3SS, invasion	*Shigella flexneri*	negative
ICE	SYI	chromosome	nitrogen fixation	*Mesorhizobium loti*	negative
cag island	PAI	chromosome	T4SS, virulence	*Helicobacter pylori*	negative

*ECI, ecological island; HPI, high pathogenicity island; ICE, integrative conjugative element; PPI-1, pneumococcal pathogenicity island 1; REI, resistance island; SAI, saprophytic island; SaPI1, *S. aureus* pathogenicity island 1 = (νSa1); SCC*mec*, staphylococcal cassette chromosome *mec*; SYI, symbiosis island.

environmental conditions. Furthermore, in all cases deletion led to the formation of non-replicative circular intermediates (Middendorf *et al.*, 2004). Deletion of PAI is not restricted to *E. coli* strain 536. The yersiniabactin island termed 'high pathogenicity island' (HPI) is relatively stable in *Y. enterocolitica*, most likely because of a mutation in the integrase gene and the lack of flanking direct repeats. In contrast, in *Y. pseudotuberculosis* excision of HPI from the chromosome occurs also by site-specific recombination between the flanking direct repeats (Bach *et al.*, 1999; Buchrieser *et al.*, 1998). Mode and frequency of deletion of HPI is again different in *Y. pestis*. There, HPI loss is part of a larger deletion of 102 kb including the adjacent pigmentation locus (*pgm*). Probably, excision is mediated by recombination between two

flanking IS*100* copies (Hare and McDonough, 1999). Although excision could be demonstrated for many GEI and this process seems to be the prerequisite for transfer of GEI by conjugation or transduction, actual horizontal transfer had been demonstrated only for SaPI1. SapPI1 is a PAI of the Gram-positive pathogen *Staphylococcus aureus*. There phage 80α is functioning as a helper phage in transduction of SaPI1 (see later discussion) (Ruzin *et al.*, 2001).

Apparently, besides gain of PAI loss of PAI also represents a 'quantum leap in evolution' and both processes have serious consequences for the microbes affected. In principle, the gain of a toxin-encoding PAI could transform a commensal into a pathogenic bacterium allowing infection of new hosts or previously unreachable niches in the same host. Later on in infection, loss of a PAI might help the microorganism to lower virulence and evade constant attack by the host's immune system. This might be a prerequisite for establishing a long-term or even chronic infection.

5.3.2. Distribution

5.3.2.1. GEI and PAI in Gram-negative Bacteria

Most PAI were identified in the genomes of Gram-negative bacteria pathogenic for humans, other animals, or plants (see Chapters 6 and 8). In the case of the prototype of such bacteria, pathogenic *Escherichia coli*, there are several PAI present in each strain. For the uropathogenic *E. coli* (UPEC) strain 536 at least five PAI and four GEI were identified (Brzuszkiewicz *et al.*, 2006). A recently discovered GEI in *E. coli* strains of phylogenetic group B2 encodes enzymes for the synthesis of peptide-polyketide hybrid compounds. This GEI termed *pks* island (54 kb) was detected in commensal (e.g., *E. coli* strain Nissle 1917) as well as extraintestinal pathogenic *E. coli* strains (e.g., the newborn meningitis strain IHE3034). Contact with *E. coli* expressing this gene cluster results in blockage of mitosis and induction of megalocytosis in mammalian cell lines and eventually leads to cell death. This seems to be the result of DNA double-strand breaks and activation of DNA damage checkpoint pathway (Nougayrede *et al.*, 2006). Genome analysis of UPEC strain (CFT073) and of an enterohemorrhagic *E. coli* (EHEC) strain of serotype O157:H7 revealed the presence of up to 13 'PAI-like' GEI per strain (Perna *et al.*, 2001). The presence of many PAI in a pathogenic *E. coli* strain seems rather to be the rule than the exception. Also, for pathogenic *Salmonella*, *Shigella*, and *Yersinia* more than one PAI could be detected per strain. The genome of *Salmonella enterica* serovar Typhimurium LT2 contains at least 15 PAI-like structures that are associated with tRNA genes, and

Table 5.2. *Comparison of typical properties of GEI in Gram-positive and Gram-negative bacteria*

	GEI properties in	
	Gram-positives	Gram-negatives
Insertion at tRNA gene	rare*	frequent
Flanked by DR	rare	frequent
Flanked by IS elements	frequent	rare
Presence of integrase/recombinase gene	rare	frequent
Different G+C content and codon usage	yes	yes
Instability	frequent	frequent

*For example, PAI I of group B *Streptococcus* strain NEM316 at tRNA-Ala gene.

S. enterica serovar Typhi acquired 10 PAI as has been revealed by analysis of the genome sequence (Parkhill *et al.*, 2001). Similarly, the *Y. pestis* genome has acquired 9 PAI-like structures (Deng *et al.*, 2002). Table 5.2 lists some of the Gram-negative bacteria for which PAI have been identified.

The ongoing sequencing projects will surely further increase the number of PAI. This illustrates the importance of PAI for the evolution of pathogenic Gram-negative bacteria. The virulence factors encoded by PAI are divergent and some PAI are characteristic for certain pathotypes (Dobrindt, 2005). In some instances the acquisition of a PAI is not sufficient for proper functioning of the encoded virulence factor. *Shigella* spp. and enteroinvasive *E. coli* harbor a large genomic deletion. This deletion resulted in loss of the *cadA* gene, which in turn caused loss of cadaverine synthesis. Complemented strains able to produce cadaverine are attenuated and enterotoxin activity is greatly inhibited (Maurelli *et al.*, 1998). This is a dramatic example of how acquisition of PAI might lead to modulation of even the core genome.

Another extreme is represented by *Pseudomonas aeruginosa*, which has the remarkable capacity to inhabit diverse environments and infect insects, plants, and animals. In this species most virulence-related genes are part of the core gene pool. Also, genes with a direct role or predicted to play a direct role in virulence are highly conserved among environmental and isolates from human infections, as in individuals with cystic fibrosis, urinary tract infections, or ocular and blood infections. However, some virulence related properties are also encoded on PAI in *P. aeruginosa*. A PAI-like 48.9-kb region (PAGI-1) in the chromosome of 85% of clinical isolates of *P. aeruginosa* from

cystic fibrosis patients is not present in the genome of strain PAO1. PAGI-1 contains IS elements and encodes regulatory proteins, a number of dehydrogenase homologs, and two proteins for detoxification of reactive oxygen species (Liang *et al.*, 2001). Another PAI, the MA island of *P. aeruginosa*, is characteristic of a highly transmissible cystic fibrosis strain. The 13-kb MA island is flanked by direct repeats, inserted close to a tRNA gene, and consists of two bacteriophage-like regions, one of which contains a *Vibrio cholerae*-like toxin gene (Lewis *et al.*, 2005). Another toxin gene present in certain *P. aeruginosa* strains is *exoU*. ExoU is a cytotoxin secreted by the same T3SS as ExoS. Interestingly, both genes, *exoU* and *exoS*, were never found together in the same strain. It is most likely that *exoU* together with its chaperon *spcU* are part of an 80-kb PAI. The acquisition of this island always leads to loss of *exoS*, which is flanked by near-perfect (9 of 10 bp) direct repeats (Wolfgang *et al.*, 2003). This is another example of the loss of genes as a consequence of the acquisition of a PAI. However, in contrast to the situation in *Shigella*, the reason for the incompatibility of *exoS* and *exoU* is unknown.

Even so, some PAI are linked to certain pathotypes in enterobacteria, whereas others can be detected in different enterobacterial species, pathotypes, or strains. The most prominent examples for the latter type of PAI is HPI, which encodes the yersiniabactin iron uptake system and the LEE (locus of enterocyte effacement) PAI. HPI is not restricted to pathogenic *Yersinia* but is widespread among enterobacteria such as *E. coli*, *Citrobacter*, *Klebsiella*, and some *Salmonella* spp. (Schubert *et al.*, 2004). However, because HPI is not present in the human pathogenic *Salmonella* group I and also found in commensal *E. coli*, it functions in these bacteria as a fitness island (Oelschlaeger *et al.*, 2003).

The LEE PAI is present in various attaching and effacing lesion-inducing enteropathogenic *E. coli* (EPEC) and EHEC strains in at least 14 different variations according to the LEE-encoded adhesin intimin. The size of LEE varies between 36 and 111 kb depending on what *E. coli* lineage LEE is carried (Jores *et al.*, 2004). Furthermore, LEE was detected in *Citrobacter rodentium* and *Hafnia alvei*. *C. rodentium* is the causative agent of murine transmissible colonic hyperplasia (Kelly *et al.*, 2006). *H. alvei* is a rare cause of bacteremia, and the role it may play in gastroenteritis is presently unknown (Janda and Abbott, 2006).

As exemplified by LEE, even PAI that seem identical at the first glance exhibit variability. This is even more pronounced for further related PAI in that they differ in composition, structural organization, and chromosomal

localization even among strains of the same patho- or serotype. PAI of animal-pathogenic, Gram-negative bacteria may encode a protein secretion system (type I, type III, or type IV), effectors translocated into host cells to subvert their function, adhesins of fimbrial or afimbrial type, toxins, iron uptake systems (e.g., the aforementioned yersiniabactin system), and capsule determinants. In addition to the diversity in the virulence function that are PAI-encoded, PAI might show a mosaic structure, especially if many IS elements are present in such PAI. These elements are suited to rearrange once-acquired PAI and add additional genes to or delete them from a certain PAI. This is not only observed for animal pathogenic bacteria but also found in plant pathogenic bacteria harboring PAI (Hochhut *et al.*, 2006a).

5.3.2.2. GEI and PAI of Gram-positive Bacteria

In Gram-positive bacteria, GEI often show properties distinct from those of GEI of Gram-negative bacteria (Table 5.2). They might lack flanking direct repeats, mobility genes, and the association with tRNA-encoding genes. However, the number of GEI reported for Gram-positive bacteria is constantly increasing.

One important example of GEI is the staphylococcal cassette chromosome *mec* (SCC*mec*) variants of *Staphylococcus aureus* (for a recent review, see Hanssen and Ericson Sollid, 2006). These GEI are responsible for the methicillin-resistant phenotype *S. aureus* strains (MRSA). Currently there are five major types of SCC*mec* described that are 21 to 67 kb in size. They are defined by the presence of *mecA*, encoding an alternate penicillin binding protein (PBP2′ or PBP2a) and its two regulatory genes *mecI* and *mecRI* as well as the variety of site-specific recombinase genes (either *ccrAB* or *ccrC*). In addition to the *mec* and *ccr* genes, SCC*mec* types II, III, and IVc contain one or more antibiotic resistance genes (Hiramatsu *et al.*, 2001). All SCC are integrated at a unique site (*attB*SCC) near the *S. aureus* origin of replication. A 15-bp sequence of *attB*SCC is found at both chromosome-SCC*mec* junctions as direct repeats. One of the two repeat sequences is located within SCC*mec* at the right end. In addition, incomplete inverted repeats are present at both ends of SCC*mec* (Ito *et al.*, 2001). These repeats are recognized by the SCC*mec*-specific recombinases during integration and excision to and from the chromosome (Hiramatsu *et al.*, 2001).

Besides SCC*mec*, there are also non-*mec* SCC. At least four non-*mec* SCC types have been described to date in *S. aureus* and coagulase-negative staphylococci (CoNS). The SCCcap1 of *S. aureus* is located also at *attB*SCC, and it contains a determinant for capsular polysaccharide 1, which makes the

strain more resistant to phagocytosis. This GEI is defective in mobilization, because it lacks a *ccrA* homolog and contains a *ccrB* homolog with a nonsense mutation (Luong *et al.*, 2002). SCC-GEI are not just responsible for horizontal spread of antibiotic resistance but might also be involved in increased virulence in *S. aureus* and CoNS. As for GEI in Gram-negatives, SCC can be stabilized by mutations in or deletions of the recombinase genes. In addition, a variety of variant SCC*mec* and non-typable SCC*mec* have been observed in different CoNS and might indicate a reservoir of SCC*mec* in CoNS. Also, the high variability of SCC and multiple copies of SCC-GEI indicate that these elements may be hot spots for recombination, a possible survival strategy for the bacteria. Furthermore, SCC can be viewed as major gene acquisition machines serving as genetic shuttles between staphylococci (Hanssen and Ericson Sollid, 2006).

Other types of GEI have been observed in *S. aureus* as well. Seven PAI (νSa) have been identified in *S. aureus* (νSa1, νSa2, νSa3, νSa4, νSaα, νSaβ, νSaγ) and one in *S. epidermidis* (νSeγ) (Gill *et al.*, 2005). These PAI carry approximately one-half of the genes for the *S. aureus* toxins or virulence factors and contribute to the pathogenic potential of this species (Gill *et al.*, 2005). Some of them are variants of or identical with earlier reported PAI such as SaPI1 (νSa1 of strain COL), SaPI2 (νSa4 of strain COL, N315, Mu50, and MW2), and SaPI3 (νSa1 of strain COL, νSa3 of strain Mu50, and MW2), to name just a few.

Various mobile genetic elements, including GEI, integrated plasmids, and prophages, make up approximately 7% of the *S. aureus* COL and 9% of the *S. epidermidis* RP62a genome. In *S. aureus* the various capabilities of virulence seem to depend on a combination of both GEI in the form of phage and PAI as well as the presence of single nucleotide polymorphism SNP. Analyses of several *S. aureus* genomes also indicate gene transfer between staphylococci and bacilli. The *cap* operon, encoding the polyglutamate capsule, has integrated in the genomes of both *S. epidermidis* RP62a and ATCC12228. This is likely the result of a plasmid-mediated gene transfer between two genera, because *cap* is encoded by a *B. anthracis* plasmid (Gill *et al.*, 2005).

Streptococcus pneumoniae is another important Gram-positive pathogen, responsible for over a million fatal infections annually (Anonymous, 2000). Certain clinical isolates of *S. pneumoniae* harbor PAI. The PPI-1 (pneumococcal pathogenicity island 1) from a serotype III clinical isolate encodes among unknown functions an iron uptake ABC transporter system (Pit2ABCD), which is important for virulence in the mouse model. Immediately downstream from *pit2ABCD* a putative recombinase gene was identified.

PPI-1 (27.2 kb) has a lower G+C content (32.6 ± 4.0%) than the mean G+C content of flanking regions (left: 40.1 + 2.1%; right: 39.4 + 4.0%) (Brown *et al.*, 2001).

Comparative genomic analysis resulted in the identification of two PAI in *S. pneumoniae* isolates from invasive pneumococcal disease. Both exhibit atypical G+C contents. PAI RD8a (ca. 20 kb) encodes an oxidoreductase, neuraminidase A, sugar-modifying enzymes, and V-type sodium synthase (Obert *et al.*, 2006). Neuraminidase A was demonstrated to contribute to pathogenesis by cleaving sialic acid residues on the surface of host cells and exposing eukaryotic receptors that enhance adhesion (Tong *et al.*, 2001). Glycosylated host components, such as secretory immunoglobulin A2, lactoferrin, and C-reactive protein, that attach to the bacteria and facilitate clearance might also be altered by the neuraminidase/sugar-modifying enzymes. PAI RD10 (36.179 kb) encodes PsrP, a protein homologous to the platelet adhesion GspB in *Streptococcus gordonii*. GspB is a sialic acid binding hemagglutinin. This adhesion mediates binding to the platelet glycoprotein Ibα and is thought to play a central role in the development of infective endocarditis (Bensing *et al.*, 2004). This PAI also encodes the export machinery, an alternate SecA/SecY transport system, most likely responsible for the secretion of PrsP. In addition, seven glycosyltransferases are encoded by RD10. The significance of this PAI was demonstrated by the fact that a RD10 deletion mutant was delayed in developing bacteremia and caused reduced mortality in the mouse model (Obert *et al.*, 2006).

Group B *Streptococcus* (GBS) is the major cause of neonatal sepsis and meningitis in the industrialized world and is an increasingly important cause of invasive disease in the elderly (Doran and Nizet, 2004). Also, this group of pathogens harbors GEI including some PAI. In GBS strain NEM316 (serotype III) 14 islands were identified, four of which might be PAI (islands I, VI, X, and XII) because they encode virulence-related genes and harbor mobilization genes (e.g., integrase genes, transposase genes). PAI I is located adjacent to the tRNA-Ala gene. Unfortunately, no information regarding G+C content of the PAI was presented in the study by Herbert *et al.* (2005).

The analysis of eight sequenced GBS genomes revealed the presence of three GEI encoding pilus-like structures. The largest of these GEI, PI-1 (16 kb), was found in six strains and is flanked by an 11-bp direct repeat. In addition to the pilus operon, PI-1 contains a gene coding for an AraC-type transcriptional regulator and homology to transposon-like genes. The other two GEI are variants of the same GEI and termed PI-2a and PI-2b. These 11-kb GEI are always inserted between genes SAG1403 and SAG1410.

Interestingly, in group A *Streptococcus* (GAS) an 11-kb region (fibronectin-binding collagen-binding T-antigen region, FCT) containing genes coding for a pilus has been found in all strains studied to date (Mora *et al.*, 2005). The organization of the FCT region is similar to that of PI-2. A similarly organized 14-kb genomic island with genes related to the GAS FCT region has been reported in *S. pneumoniae* and has been shown to code for pilus-like structures (Barocchi *et al.*, 2006). This 14-kb GEI of *S. pneumoniae* is flanked by IS*1167* (Hava and Camilli, 2002). It has been suggested that the GAS and *S. pneumoniae* pilus GEI derive from a common ancestor and have been acquired by horizontal gene transfer (Bessen and Kalia, 2002). All these pilus-like structures tested so far are able to induce protective immunity in the mouse models of both GBS and GAS disease (Mora *et al.*, 2005).

The genome of GAS *Streptococcus* strain MGAS6180 (serotype M28) contains a 37.4-kb GEI designated 'region of difference 2' (RD2). RD2 is similar in gene content and organization to regions described as GEI in serotype II and V GBS strains NEM316 and 2603V/R. GEI RD2 has multiple genes with orthologs in prophages and plasmids; its GC content (35.1%) is lower than the average for GAS (38.3%) and is close to that of GBS (35.7%). This suggests that this GEI was acquired by horizontal gene transfer. Moreover, similar GEI were identified in six other GBS strains (serotypes Ia, Ib, II, III, and V) (Tettelin *et al.*, 2005). RD2 of MGAS6180 encodes seven proteins with Gram-positive secretion signal sequences. One of them is a member of the antigen I/II family of surface-anchored molecules produced by oral streptococci (Zhang *et al.*, 2006). These antigens are adhesions that bind human salivary glycoproteins and assist colonization of the oropharynx (Jakubovics *et al.*, 2005).

Streptococcus mutans is implicated as the principal causative agent of human dental caries. Analysis of the complete genome revealed the presence of least nine GEI in the genome of *S. mutans* (Waterhouse *et al.*, 2007). These GEI can be lost and are therefore not present in all *S. mutans* strains investigated. They encode bacteriocins, parts of the histidine metabolism, bacitracin synthetases, DNA modifying enzymes, and often transposases. Some of these GEI have comparable analogs in *S. pneumoniae* or seem to originate from a *Bacillus* species.

Similar to the Ipa-PAI of *Shigella*, a 37-kb PAI is located on a plasmid, pOX2, of *Bacillus anthracis*. This PAI contains genes of pOX2 all known to play a capital role in the course of anthrax infection, for example, the capsule genes. This island also has a lower G+C content of 30.9% versus 34.3% for the core genome and is rich in mobile genetic elements such as multiple inactivated copies of the IS*231* transposase and a vestigial class II transposon.

Furthermore, all the capsule genes and their associated regulatory elements can be considered an IS*231*-derived mobile insertion cassette (MIC) flanked by IS*231*-related inverted repeats (Van der Auwera *et al.*, 2005). This is paralleled in pXO1 of *B. anthracis* by the presence of a class II transposon, Tn*XO1*, in the PAI of this plasmid. Tn*XO1* carries the anthrax germination operon *gerX* as its passenger genes (Van der Auwera and Mahillon, 2005). The pXO1 PAI (44.8 kb) is flanked by two almost identical copies of the IS*1627* element, contains besides the already mentioned *gerX* genes all the known toxin genes (*cya*, *lef*, and *pagA*) and toxin regulatory elements (*atxA* and *pagR*), several putative IS elements, and genes encoding integrase and transposases (Okinaka *et al.*, 1999). Evidence for the mobility of this PAI is its inverted orientation on pXO1 plasmids originating from two different strains of *B. anthracis* (Thorne, 1993).

An example of a GEI encoding an important metabolic property of a pathogenic species is the ethanolamine degradation encoding island present in most *Clostridium difficile* strains. The use of ethanolamine might be important for the anaerobic intestinal lifestyle of *C. difficile*, since ethanolamine is a carbon and nitrogen source provided by the host's dietary intake (Stabler *et al.*, 2006). This GEI is inserted into gene CD1927 and the last gene of this island encodes a site-specific recombinase, typical for GEI.

The virulence of another Gram-positive enteropathogen is clearly dependent on the presence of a certain PAI. This is the 6036-bp *Bacteroides fragilis* pathogenicity island (BfPAI), which encodes a zinc metalloprotease toxin. BfPAI is inserted into the genome of *B. fragilis* between the stop codons of *bfmB* (mobilization gene) and *bfmC* (a *traD* homolog) and has a much lower C+G content (35%) than the core genome (42%) (Franco *et al.*, 1999).

5.4. Tn*7* AS A FOUNDER MODEL FOR GEI

Tn*7* is an example of GEI formation in Gram-positive and Gram-negative bacteria (Parks and Peters, 2007). Tn*7* encodes five proteins (TnsA, TnsB, TnsC, TnsD, and TnsE) essential for two transposition pathways. The TnsAbc proteins constitute the core transposition machinery that interacts with one of the two target site-selecting proteins, TnsD or TnsE. TnsD-mediated transposition is directed into the *att*Tn*7* site. This site is located in the transcriptional terminator for the *glmS* gene. In contrast, TnsE does not recognize any particular DNA sequence but preferentially directs transposition into mobile plasmids. These properties allowed Tn*7* and relatives thereof to become widespread in disparate environments and phylogenetically

diverse species. Furthermore, Tn*7* seems to be able to initiate genomic island formation by functioning as a founder element that brings with it other mobile elements and possibly attachment sites for bacteriophages, integrons, or other transposons. This is best exemplified by various Tn*7* derivates and even phage-related genes all located in the *att*Tn*7* site of several *Shewanella* species. These islands might reflect stages of the evolution from Tn*7* to various GEI, which might occur in Gram-positive as well as in Gram-negative bacteria, because Tn*7* derivates are found in both groups of microorganisms (Parks and Peters, 2007).

5.5. GEI IN BACTERIAL EVOLUTION

For the past few decades bacterial evolution had been seen as a rather slow process based in principle on point mutations, deletions, duplications, and inversions of DNA already present in the bacterial cell. The results of these genetic changes were inherited from one generation to the next. Now we know that exchange of genetic material in the form of whole functional determinants between microorganisms of one generation is achieved frequently on the evolutionary scale. This horizontal transfer of genes by conjugation mediated by plasmids or transfection via bacteriophages and even by transformation with 'naked' DNA became obvious and was appreciated as a means of rapid evolution. Some of these horizontally acquired determinants are GEI.

The acquisition of a PAI can be sufficient to transform a less aggressive pathogen in a more virulent one, as in the case of the presence of the *cag* island of *Helicobacter pylori* strains of type I, which are associated with severe forms of gastroduodenal disease (Censini *et al.*, 1996). Similarly, Gram-positive *Enterococcus faecalis* strains causing more severe infections harbor a chromosome- or plasmid-located PAI encoding a cytolysin, in contrast to strains that are associated with less severe infections (Shankar *et al.*, 2004). Furthermore, cloning of the LEE, a PAI first described in a diarrheagenic EPEC strain, into an apathogenic *E. coli* K-12 strain confers the complete attaching and effacing phenotype of the EPEC strain. This is a striking example that avirulent bacteria can be transformed into pathogenic ones through a single genetic step (McDaniel and Kaper, 1997).

The assumption that GEI can be laterally spread by phages was best demonstrated for the SaPI1 PAI of *S. aureus*, which can be excised, brought into a circular form, and induced to replicate by the helper bacteriophage 80α. Furthermore, the excised PAI is transduced to other strains with rather

high frequency where SaPI1 integrates at the same site of the genome, that is, at *attC* by means of a self-coded integrase (Ruzin *et al.*, 2001).

Most important, the acquisition of a GEI results in assimilation of large numbers of genes, often of entire operons that confer new traits. This horizontal transfer into the recipient's genome results in dramatic changes of properties. The new properties might provide a selective advantage under certain conditions, leading to enhanced transmission and improved colonization and resulting in the ability to occupy new niches in the common host or even new host organisms. Obviously, GEI that encode genes which increase the ability of the bacterial recipient to adapt underlie positive selection. Once a GEI has been integrated into the recipient's genome, it may be modulated by recombination and deletion events, resulting, for example, in inactivation or loss of mobility genes, and therefore becomes a stable part of the genome (Dobrindt *et al.*, 2002). Selection can even result in loss or inactivation of core genes, the presence or activity of which might interfere with the function of the GEI encoded genes. This is exemplified by the already mentioned deletion of the *cadA* gene in virulent *Shigella*.

There is also collaboration among different PAI as well as between PAI and other mobile genetic elements as well as core genome encoded functions. For EHEC as well as EPEC a major PAI was identified and termed LEE. LEE encodes a T3SS that translocates not only effectors encoded by LEE but several other effectors encoded either by prophages (e.g., EspI = NleA, EspJ, EspFu = TccP) or other PAI. The effector EspG2 is encoded by the EspC-PAI and NleB and NleE by a PAI known as O-island 122 (Garmendia *et al.*, 2005). Similarly, in *Salmonella* SPI-1 encodes a T3SS that delivers effectors into eukaryotic host cells. One of these effectors is SopE not encoded by SPI-1 but by a prophage (Mirold *et al.*, 1999).

The sources for GEI are most likely niches harboring diverse microbial communities such as mixed biofilms or the rumen and guts of animals. However, there are indications that besides prokaryotic genes there are also genes presumably acquired by horizontal gene transfer from a eukaryotic origin. This might explain the presence of numerous eukaryotic-like proteins in *Legionella* spp. (Cazalet *et al.*, 2004).

The integration of newly acquired GEI results not just in new GEI-encoded properties but also in the establishment of new regulatory and functional networks, which is the case for the regulation of LEE gene expression by regulators encoded within this PAI, on a plasmid and on the core genome (Deng *et al.*, 2004). This process becomes even more complex by lateral acquisition of further genes and rearrangement, inactivation, or loss of other core genome or GEI-encoded determinants. Recently, the nucleoid

protein H-NS has been shown to selectively silence horizontally acquired genes in *Salmonella* by targeting sequences with G+C content lower than that of the core genome, including SPI-2, SPI-3, and SPI-5 (Navarre *et al.*, 2005). All this allows selective forces to fine tune GEI-encoded functions in the context of core genome properties. In summary, horizontal transfer of GEI is a major and continuing force in the fast stepwise evolution of new pathotypes or even new pathogens by so-called evolution in quantum leaps and for the spread of antibiotic resistance.

REFERENCES

Anonymous (2000). Preventing pneumococcal disease among infants and young children. Recommendations of the Advisory Committee on Immunization Practices (ACIP). *Morb Mortal Wkly Rep Recomm Rep*, **49**, 1–35.

Bach, S., Buchrieser, C., Prentice, M., *et al.* (1999). The high-pathogenicity island of *Yersinia enterocolitica* Ye8081 undergoes low-frequency deletion but not precise excision, suggesting recent stabilization in the genome. *Infect Immun*, **67**, 5091–9.

Barocchi, M. A., Ries, J., Zogaj, X., *et al.* (2006). A pneumococcal pilus influences virulence and host inflammatory responses. *Proc Natl Acad Sci USA*, **103**, 2857–62.

Bennett, P. M. (2004). Genome plasticity: insertion sequence elements, transposons and integrons, and DNA rearrangement. *Methods Mol Biol*, **266**, 71–113.

Bensing, B. A., Lopez, J. A., and Sullam, P. M. (2004). The *Streptococcus gordonii* surface proteins GspB and Hsa mediate binding to sialylated carbohydrate epitopes on the platelet membrane glycoprotein Ibα. *Infect Immun*, **72**, 6528–37.

Bessen, D. E., and Kalia, A. (2002). Genomic localization of a T serotype locus to a recombinatorial zone encoding extracellular matrix-binding proteins in *Streptococcus pyogenes*. *Infect Immun*, **70**, 1159–67.

Brown, J. S., Gilliland, S. M., and Holden, D. W. (2001). A *Streptococcus pneumoniae* pathogenicity island encoding an ABC transporter involved in iron uptake and virulence. *Mol Microbiol*, **40**, 572–85.

Brüssow, H., Canchaya, C., and Hardt, W. D. (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol Rev*, **68**, 560–602.

Brzuszkiewicz, E., Brüggemann, H., Liesegang, H., *et al.* (2006). How to become a uropathogen: comparative genomic analysis of extraintestinal pathogenic *Escherichia coli* strains. *Proc Natl Acad Sci USA*, **103**, 12879–84.

Buchrieser, C., Brosch, R., Bach, S., Guiyoule, A., and Carniel, E. (1998). The high-pathogenicity island of *Yersinia pseudotuberculosis* can be inserted into any of the three chromosomal *asn* tRNA genes. *Mol Microbiol*, **30**, 965–78.

Cazalet, C., Rusniok, C., Brüggemann, H., *et al.* (2004). Evidence in the *Legionella pneumophila* genome for exploitation of host cell functions and high genome plasticity. *Nat Genet*, **36**, 1165–73.

Censini, S., Lange, C., Xiang, Z., *et al.* (1996). *cag*, a pathogenicity island of *Helicobacter pylori*, encodes type I-specific and disease-associated virulence factors. *Proc Natl Acad Sci USA*, **93**, 14648–53.

Collyn, F., Billault, A., Mullet, C., Simonet, M., and Marceau, M. (2004). YAPI, a new *Yersinia pseudotuberculosis* pathogenicity island. *Infect Immun*, **72**, 4784–90.

Deng, W., Burland, V., Plunkett, G. 3rd, *et al.* (2002). Genome sequence of *Yersinia pestis* KIM. *J Bacteriol*, **184**, 4601–11.

Deng, W., Puente, J. L., Gruenheid, S., *et al.* (2004). Dissecting virulence: systematic and functional analyses of a pathogenicity island. *Proc Natl Acad Sci USA*, **101**, 3597–602.

Dobrindt, U. (2005). (Patho-)Genomics of *Escherichia coli*. *Int J Med Microbiol*, **295**, 357–71.

Dobrindt, U., Hentschel, U., Kaper, J. B., and Hacker, J. (2002). Genome plasticity in pathogenic and nonpathogenic enterobacteria. *Curr Top Microbiol Immunol*, **264**, 157–75.

Dobrindt, U., Hochhut, B., Hentschel, U., and Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol*, **2**, 414–24.

Doran, K. S., and Nizet, V. (2004). Molecular pathogenesis of neonatal group B streptococcal infection: no longer in its infancy. *Mol Microbiol*, **54**, 23–31.

Franco, A. A., Cheng, R. K., Chung, G. T., *et al.* (1999). Molecular evolution of the pathogenicity island of enterotoxigenic *Bacteroides fragilis* strains. *J Bacteriol*, **181**, 6623–33.

Garmendia, J., Frankel, G., and Crepin, V. F. (2005). Enteropathogenic and enterohemorrhagic *Escherichia coli* infections: translocation, translocation, translocation. *Infect Immun*, **73**, 2573–85.

Germon, P., Roche, D., Melo, S., *et al.* (2007). tDNA locus polymorphism and ecto-chromosomal DNA insertion hot-spots are related to the phylogenetic group of *Escherichia coli* strains. *Microbiology*, **153**, 826–37.

Gill, S. R., Fouts, D. E., Archer, G. L., *et al.* (2005). Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant *Staphylococcus aureus* strain and a biofilm-producing

methicillin-resistant *Staphylococcus epidermidis* strain. *J Bacteriol*, **187**, 2426–38.

Groisman, E. A., and Ochman, H. (1996). Pathogenicity islands: bacterial evolution in quantum leaps. *Cell*, **87**, 791–4.

Hacker, J., and Carniel, E. (2001). Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep*, **2**, 376–81.

Hacker, J., and Kaper, J. B. (1999). The concept of pathogenicity islands. In Hacker, J., and Kaper, J. B. (Eds.). *Pathogenicity islands and other mobile virulence elements*. Washington, DC: American Society for Microbiology.

Hacker, J., and Kaper, J. B. (2000). Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol*, **54**, 641–79.

Hacker, J., Bender, L., Ott, M., *et al.* (1990). Deletions of chromosomal regions coding for fimbriae and hemolysins occur *in vitro* and *in vivo* in various extraintestinal *Escherichia coli* isolates. *Microb Pathog*, **8**, 213–25.

Hall, R. M., and Collis, C. M. (1995). Mobile gene cassettes and integrons: capture and spread of genes by site-specific recombination. *Mol Microbiol*, **15**, 593–600.

Hanssen, A. M., and Ericson Sollid, J. U. (2006). SCC*mec* in staphylococci: genes on the move. *FEMS Immunol Med Microbiol*, **46**, 8–20.

Hare, J. M., and McDonough, K. A. (1999). High-frequency RecA-dependent and -independent mechanisms of Congo red binding mutations in *Yersinia pestis*. *J Bacteriol*, **181**, 4896–904.

Hava, D. L., and Camilli, A. (2002). Large-scale identification of serotype 4 *Streptococcus pneumoniae* virulence factors. *Mol Microbiol*, **45**, 1389–406.

Herbert, M. A., Beveridge, C. J., McCormick, D., *et al.* (2005). Genetic islands of *Streptococcus agalactiae* strains NEM316 and 2603VR and their presence in other Group B streptococcal strains. *BMC Microbiol*, **5**, 31.

Hiramatsu, K., Cui, L., Kuroda, M., and Ito, T. (2001). The emergence and evolution of methicillin-resistant *Staphylococcus aureus*. *Trends Microbiol*, **9**, 486–93.

Hochhut, B., Dobrindt, U., and Hacker, J. (2006a). The contribution of pathogenicity islands to the evolution of bacterial pathogens. In Seifert, H. S., and Dirita, V. J. (Eds.). *Evolution of microbial pathogens*. Washington, DC: American Society for Microbiology.

Hochhut, B., Wilde, C., Balling, G., *et al.* (2006b). Role of pathogenicity island-associated integrases in the genome plasticity of uropathogenic *Escherichia coli* strain 536. *Mol Microbiol*, **61**, 584–95.

Ito, T., Katayama, Y., Asada, K., *et al.* (2001). Structural comparison of three types of staphylococcal cassette chromosome *mec* integrated in the chromosome

in methicillin-resistant *Staphylococcus aureus*. *Antimicrob Agents Chemother*, **45**, 1323–36.

Jakubovics, N. S., Stromberg, N., Van Dolleweerd, C. J., Kelly, C. G., and Jenkinson, H. F. (2005). Differential binding specificities of oral streptococcal antigen I/II family adhesins for human or bacterial ligands. *Mol Microbiol*, **55**, 1591–605.

Janda, J. M., and Abbott, S. L. (2006). The genus *Hafnia*: from soup to nuts. *Clin Microbiol Rev*, **19**, 12–8.

Jores, J., Rumer, L., and Wieler, L. H. (2004). Impact of the locus of enterocyte effacement pathogenicity island on the evolution of pathogenic *Escherichia coli*. *Int J Med Microbiol*, **294**, 103–13.

Kelly, M., Hart, E., Mundy, R., *et al.* (2006). Essential role of the type III secretion system effector NleB in colonization of mice by *Citrobacter rodentium*. *Infect Immun*, **74**, 2328–37.

Lewis, D. A., Jones, A., Parkhill, J., *et al.* (2005). Identification of DNA markers for a transmissible *Pseudomonas aeruginosa* cystic fibrosis strain. *Am J Respir Cell Mol Biol*, **33**, 56–64.

Liang, X., Pham, X. Q., Olson, M. V., and Lory, S. (2001). Identification of a genomic island present in the majority of pathogenic isolates of *Pseudomonas aeruginosa*. *J Bacteriol*, **183**, 843–53.

Luong, T. T., Ouyang, S., Bush, K., and Lee, C. Y. (2002). Type 1 capsule genes of *Staphylococcus aureus* are carried in a staphylococcal cassette chromosome genetic element. *J Bacteriol*, **184**, 3623–9.

Maurelli, A. T., Fernandez, R. E., Bloch, C. A., Rode, C. K., and Fasano, A. (1998). "Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. *Proc Natl Acad Sci USA*, **95**, 3943–8.

Mazel, D. (2006). Integrons: agents of bacterial evolution. *Nat Rev Microbiol*, **4**, 608–20.

McDaniel, T. K., and Kaper, J. B. (1997). A cloned pathogenicity island from enteropathogenic *Escherichia coli* confers the attaching and effacing phenotype on *E. coli* K-12. *Mol Microbiol*, **23**, 399–407.

Middendorf, B., Hochhut, B., Leipold, K., *et al.* (2004). Instability of pathogenicity islands in uropathogenic *Escherichia coli* 536. *J Bacteriol*, **186**, 3086–96.

Mirold, S., Rabsch, W., Rohde, M., *et al.* (1999). Isolation of a temperate bacteriophage encoding the type III effector protein SopE from an epidemic *Salmonella typhimurium* strain. *Proc Natl Acad Sci USA*, **96**, 9845–50.

Mongkolrattanothai, K., Boyle, S., Murphy, T. V., and Daum, R. S. (2004). Novel non-*mecA*-containing staphylococcal chromosomal cassette composite island containing *pbp4* and *tagF* genes in a commensal staphylococcal

species: a possible reservoir for antibiotic resistance islands in *Staphylococcus aureus*. *Antimicrob Agents Chemother*, **48**, 1823–36.

Mora, M., Bensi, G., Capo, S., *et al.* (2005). Group A *Streptococcus* produce pilus-like structures containing protective antigens and Lancefield T antigens. *Proc Natl Acad Sci USA*, **102**, 15641–6.

Nandi, S., Maurer, J. J., Hofacre, C., and Summers, A. O. (2004). Gram-positive bacteria are a major reservoir of Class 1 antibiotic resistance integrons in poultry litter. *Proc Natl Acad Sci USA*, **101**, 7118–22.

Navarre, W. W., Halsey, T. A., Walthers, D., *et al.* (2005). Co-regulation of *Salmonella enterica* genes required for virulence and resistance to antimicrobial peptides by SlyA and PhoP/PhoQ. *Mol Microbiol*, **56**, 492–508.

Nougayrede, J. P., Homburg, S., Taieb, F., *et al.* (2006). *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells. *Science*, **313**, 848–51.

Obert, C., Sublett, J., Kaushal, D., *et al.* (2006). Identification of a candidate *Streptococcus pneumoniae* core genome and regions of diversity correlated with invasive pneumococcal disease. *Infect Immun*, **74**, 4766–77.

Oelschlaeger, T. A., Zhang, D., Schubert, S., *et al.* (2003). The high-pathogenicity island is absent in human pathogens of *Salmonella enterica* subspecies I but present in isolates of subspecies III and VI. *J Bacteriol*, **185**, 1107–11.

Okinaka, R. T., Cloud, K., Hampton, O., *et al.* (1999). Sequence and organization of pXO1, the large *Bacillus anthracis* plasmid harboring the anthrax toxin genes. *J Bacteriol*, **181**, 6509–15.

Parkhill, J., Dougan, G., James, K. D., *et al.* (2001). Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature*, **413**, 848–52.

Parks, A. R., and Peters, J. E. (2007). Transposon Tn*7* is widespread in diverse bacteria and forms genomic islands. *J Bacteriol*, **189**, 2170–3.

Perna, N. T., Plunkett, G. 3rd, Burland, V., *et al.* (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*, **409**, 529–33.

Ruzin, A., Lindsay, J., and Novick, R. P. (2001). Molecular genetics of SaPI1 – a mobile pathogenicity island in *Staphylococcus aureus*. *Mol Microbiol*, **41**, 365–77.

Schubert, S., Rakin, A., and Heesemann, J. (2004). The *Yersinia* high-pathogenicity island (HPI): evolutionary and functional aspects. *Int J Med Microbiol*, **294**, 83–94.

Shankar, N., Coburn, P., Pillar, C., Haas, W., and Gilmore, M. (2004). Enterococcal cytolysin: activities and association with other virulence traits in a pathogenicity island. *Int J Med Microbiol*, **293**, 609–18.

Stabler, R. A., Gerding, D. N., Songer, J. G., *et al.* (2006). Comparative phylogenomics of *Clostridium difficile* reveals clade specificity and microevolution of hypervirulent strains. *J Bacteriol*, **188**, 7297–305.

Tettelin, H., Masignani, V., Cieslewicz, M. J., *et al.* (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome." *Proc Natl Acad Sci USA*, **102**, 13950–5.

Thorne, C. (1993). *Bacillus anthracis*. In Soneshein, A. L. (Ed.). *Bacillus subtilis and other gram-positive bacteria*. Washington, DC: American Society for Microbiology.

Tong, H. H., James, M., Grants, I., *et al.* (2001). Comparison of structural changes of cell surface carbohydrates in the eustachian tube epithelium of chinchillas infected with a *Streptococcus pneumoniae* neuraminidase-deficient mutant or its isogenic parent strain. *Microb Pathog*, **31**, 309–17.

Van der Auwera, G. A., Andrup, L., and Mahillon, J. (2005). Conjugative plasmid pAW63 brings new insights into the genesis of the *Bacillus anthracis* virulence plasmid pXO2 and of the *Bacillus thuringiensis* plasmid pBT9727. *BMC Genomics* **6**, 103–110.

Van der Auwera, G.A., and Mahillon, J. (2005). TnXO1, a germination-associated class II transposon from *Bacillus anthracis*. *Plasmid*, **53**, 251–7.

Waterhouse, J. C., Swan, D. C., and Russell, R. R. (2007). Comparative genome hybridization of *Streptococcus mutans* strains. *Oral Microbiol Immunol*, **22**, 103–10.

Wolfgang, M. C., Kulasekara, B. R., Liang, X., *et al.* (2003). Conservation of genome content and virulence determinants among clinical and environmental isolates of *Pseudomonas aeruginosa*. *Proc Natl Acad Sci USA*, **100**, 8484–9.

Zhang, S., Green, N. M., Sitkiewicz, I., Lefebvre, R. B., and Musser, J. M. (2006). Identification and characterization of an antigen I/II family protein produced by group A *Streptococcus*. *Infect Immun*, **74**, 4200–13.

Part III Paradigms of Bacterial Evolution

CHAPTER 6

Genomic Islands in Plant-pathogenic Bacteria

Dawn L. Arnold and Robert W. Jackson

6.1. INTRODUCTION

The evolution of bacterial pathogens from non-pathogens or from avirulent strains is a major cause for concern in agriculture. As exemplified by the explosion of antibiotic resistance in human pathogens, bacteria can rapidly overcome control strategies and host resistance. As we are now discovering, the intrinsic plasticity of the bacterial genome combined with horizontal gene transfer is the major determinant influencing the expression of pathogenicity. Many of the disease symptoms caused by pathogens on plants, including blights, galls, chlorosis, scabs, leaf spots, and wilting, are attributable to genes that are often clustered together and, in some cases, acquired from distantly related bacteria. One mechanism that affects the virulence of plant pathogens is the loss or gain of DNA regions called genomic islands (GEI).

GEI were first described as pathogenicity islands (PAI) in human pathogenic *Escherichia coli* by Hacker *et al.* (1990), who discovered that a region of chromosomally located, virulence-associated genes of uropathogenic *E. coli* was absent from some *E. coli* isolates (Blum *et al.*, 1994). The term GEI is now more appropriate given that the features of PAI are displayed by a number of regions of DNA with functions other than pathogenicity, for example, symbiosis, metabolic, or resistance islands (Hacker and Kaper, 2000; Hentschel and Hacker, 2001). The general features of GEI are as follows: (1) They are present in the genomes of some strains but not others; (2) they carry genes of importance to the bacterial host, such as virulence or symbiosis genes; (3) they often have an altered G+C content compared to the rest of the genome, which has been suggested as an indicator of horizontal transfer; (4) they are flanked by specific DNA sequences, such as direct repeats or tRNA loci; and (5) they carry genes coding for genetic mobility

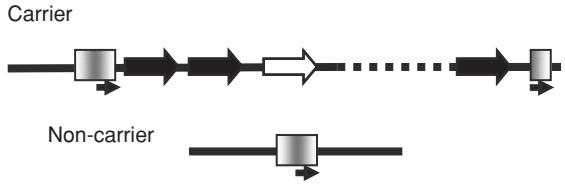


Figure 6.1. The basic structure of a genomic island (GEI). In a strain lacking a GEI (non-carrier), the box represents a target site (*att*), such as a tRNA gene or transposase, in the genome for the integration of a DNA element containing DNA homologous to the *att* target site. Integration leads to partial duplication (direct repeats) of *att* on either side of the island (arrows). The red arrow represents a gene providing a new function to the host bacterium, such as a pathogenicity or toxin gene cluster. Blue arrows represent conserved genes including insertion sequence (IS) elements, integrases, and transposases.

such as phage genes, insertion sequence elements, integrases, transposases, and origins of replication (Hacker and Kaper, 2000) (Figure 6.1).

We review the range of GEI described in plant pathogens and highlight how the explosion in available genome sequences of plant-pathogenic bacteria has led to the prediction of many more GEI. We also look at the evidence for mobility of these GEI and highlight insights that can be gained by looking at the mechanisms of mobility seen in medically important bacteria.

6.2. TYPE III SECRETION ISLANDS

Type III protein secretion systems (T3SS) are utilized by many plant- and animal-associated bacteria to promote pathogenesis or symbiosis. Several genes encode proteins that form a pore complex spanning the bacterial inner and outer membranes and an extended pilus for the secretion of proteins. In plant pathogenic bacteria, the T3SS of *Erwinia, Pseudomonas, Xanthomonas,* and *Ralstonia* are encoded by *hrp/hrc* (hypersensitive response and pathogenicity/conserved) genes (Alfano and Collmer, 1997). The T3SS is used for translocation of proteins from the bacterial cell into the plant cell (Casper-Lindley *et al.*, 2002). If a bacterial gene or cognate secreted protein triggers a defensive reaction in the plant that leads to a resistance response termed the hypersensitive reaction (HR), then it is termed an avirulence gene/protein (Keen, 1990). Alternatively, they are designated virulence genes or proteins if they lead to the development of disease symptoms in the plant. The term 'effector' has been adopted to encompass genes whose protein products contribute to the plant-microbe interaction, irrespective of function as avirulence or virulence factors (van Dijk *et al.*, 1999).

On the basis of their distinct gene arrangements and regulatory components, the *hrp/hrc* gene clusters of four plant pathogenic bacteria can be divided into two groups: I (*Pseudomonas* and *Erwinia*) and II (*Xanthomonas* and *Ralstonia*) (Alfano and Collmer, 1997). The discrepancy between the distribution of these groups and the phylogeny of the bacteria provides some evidence that *hrp/hrc* gene clusters have been horizontally acquired and therefore may represent PAI. The observation that the cluster is inserted into a tRNA in the ancestral genome supports this claim (Alfano *et al.*, 2000).

6.2.1. *Pseudomonas syringae*

The *P. syringae hrp* PAI has a tripartite mosaic structure composed of an exchangeable effector locus (EEL), the *hrp/hrc* gene cluster, and a conserved effector locus (CEL; Alfano *et al.*, 2000; Figure 6.2a). The EEL harbors exchangeable effector genes and makes only a quantitative contribution to parasitic fitness in host plants whereas the CEL contains a number of conserved open reading frames (ORF) and is essential for *P. syringae* pv. *tomato* pathogenicity in tomato. Two other studies have sequenced EEL from a range of *P. syringae* pathovars (Deng *et al.*, 2003; Charity *et al.*, 2003). These studies suggested that the EEL effector genes were acquired by horizontal transfer after the acquisition of the *hrp/hrc* gene cluster, but before the divergence of modern pathovars.

A recent study has shown that not all *Pseudomonas hrp* PAI have a tripartite structure. Araki *et al.* (2006) examined the structure and sequence of *hrp* PAI in *P. viridiflava* strains that were collected from naturally occurring *Arabidopsis thaliana* plants. They found that two structurally distinct and highly diverged PAI paralogues (T- and S-PAI) were integrated in different chromosome locations and that only one of the PAIs was ever present in any individual cell. The T-PAI was tripartite in structure with a central *hrp/hrc* cluster flanked by an EEL and a CEL, but the S-PAI encoded only the *hrp/hrc* cluster. Pathogenicity tests revealed that T-PAI isolates elicit an HR in *A. thaliana* Col-0 significantly more slowly than do S-PAI isolates. However, in tobacco, the opposite pattern was observed. This raises the possibility that the two distinct PAI are maintained by selection within the population as alternative means of interacting with different hosts.

6.2.2. *Erwinia/Pantoea*

Erwinia amylovora causes fire blight, a devastating disease of rosaceous plants. To infect plants successfully, *E. amylovora* requires products of

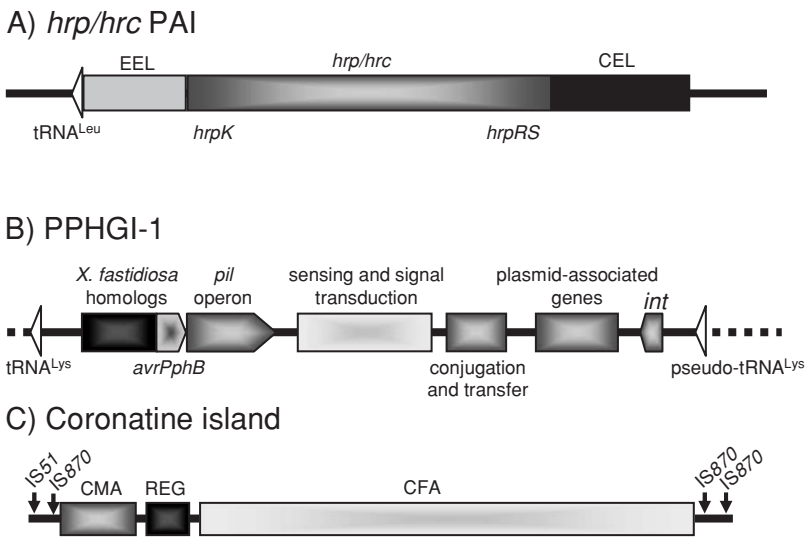


Figure 6.2. Types of genomic island found in plant-pathogenic bacteria. (A) The tripartite structure of the Hrp pathogenicity island of *P. syringae* pv. *tomato* DC3000. The core T3SS structural gene cluster is surrounded by regions that carry clusters of effector genes: EEL, exchangeable effector locus; CEL, conserved effector locus (Alfano *et al.*, 2000). (B) PPHGI-1 from *P. syringae* pv. *phaseolicola* 1302A carries the effector gene *avrPphB*. The GEI is bounded by a $tRNA^{Lys}$ gene/pseudogene and contains an integrase gene (*xerC*) just inside the right end. Colored blocks represent clusters of genes of different functional groups (Pitman *et al.*, 2005). (C) The coronatine (COR) phytotoxin gene cluster from *P. syringae* pv. *glycinea* PG4180. The cluster represents 32.8 kb of a 90-kb plasmid and is bounded by insertion sequences (IS) (Alarcón-Chaidez *et al.*, 1999). Shaded blocks represent clusters of genes: CMA, coronamic acid synthesis; CFA, coronafacic acid synthesis; REG, regulatory. Individual genes are not shown in any diagram.

juxtaposed *hrp* and disease-specific (*dspA/E* and *dspB/F*) genes (Bogdanove *et al.*, 1998). Sequence analysis of the region bordering the *hrp/dsp* gene cluster of *E. amylovora* strain Ea321 revealed characteristics of PAIs and was designated the *hrp* PAI of *E. amylovora* (Oh *et al.*, 2005). It comprises a 62-kb chromosomal region with 60 ORF that include the *hrp/dsp* genes, a phage integrase, a $tRNA^{Phe}$, several orthologues of the *Yersinia pseudotuberculosis* PAI genes, and several putative virulence genes. It was also revealed that the cluster was interrupted by four genes, only one of which, a lytic transglycosylase, was implicated in T3SS function.

The *hrp* gene cluster of *Pantoea stewartii* ssp. *stewartii* exhibits strong synteny to the *E. amylovora hrp* gene cluster, with a pair of *dspA/E* and *dspB/F* homologues, designated *wtsEF*, located next to the cluster; these genes are essential for pathogenicity to corn (Frederick *et al.*, 2001). *E. carotovora* subspecies *atroseptica* carries both *hecAB* and *dspAB*(*EF*) on one side of the *hrp* cluster, but no *plcA* (Bell *et al.*, 2002; Toth *et al.*, 2003).

E. herbicola pv. *gypsophilae* induces gall formation on gypsophila, whereas *E. herbicola* pv. *betae* elicits galls on beet. The pathogenicity of each pathovar is dependent on the presence of a similar indigenous pathogenicity plasmid $pPATH_{Pag}$ and $pPATH_{Pab}$. A 55-kb PAI, comprising the *hrp* genes and flanked by *dspAB/EF* on one side and $avrPphD_{Ehg}$, *hsvG*, cytokinin biosynthesis, and indole-3-acetic acid biosynthesis genes on the opposite side, is located on the 150-kb $pPATH_{Pag}$ plasmid. All genes contribute to the establishment of galls on *Gypsophila paniculata* stems (Mor *et al.*, 2001; Guo *et al.*, 2002).

6.2.3. *Xanthomonas*

In *Xanthomonas campestris* pv. *vesicatoria*, the T3SS *hrp* cluster is flanked by insertion sequences and a $tRNA^{Arg}$ gene (Noël *et al.*, 2002). On one side of the cluster are putative type III effector genes, *xopA* and *xopD* (*Xanthomonas* outer proteins) and an *hrpG*-induced gene, *hpaH*. On the opposite side of the *hrp* cluster are the genes *hpaF*, *hpaG*, and *hrpF* (Büttner and Bonas, 2002; Noël *et al.*, 2002; Rossier *et al.*, 2000). The region of the *hrpF* gene of *X. oryzae* pv. *oryzae* is bounded by two IS elements (Sugio *et al.*, 2005). A comparison of the *hrpF* genes of different species of *Xanthomonas* revealed that the *hrpF* region is a stable yet more variable peninsula of the *hrp* PAI. Another study compared the *hrp* PAI from *X. axonopodis* pv. *glycines* to the PAIs of four other *Xanthomonas* species (Kim *et al.*, 2003). They found that the core of the *hrp* cluster was highly conserved (>90% similarity), but the flanking regions were relatively diverse, and they concluded that, in contrast to the *P. syringae hrp* PAI, the concept of tripartite mosaic architecture was not applicable to a xanthomonad *hrp* PAI.

6.2.4. *Ralstonia*

In contrast to the other *hrp* PAIs, the 23 kb *Ralstonia hrp* cluster is located on a megaplasmid but does not have the hallmarks of a GEI, possibly indicating it to be more ancient and stabilized (Genin and Boucher, 2004).

6.3. NON-*hrp/hrc* GEI ENCODING EFFECTOR GENES

A number of effector genes, often associated with mobile elements (Kim *et al.*, 1998), have been reported on PAI not linked to the T3SS gene cluster.

6.3.1. PPHGI-1 ICEland

The *P. syringae* pv. *phaseolicola* effector *avrPphB* was initially located within a putative GEI inserted in an *att* sequence (a $tRNA^{Lys}$ gene on one side and $tRNA^{Lys}$ pseudogene on the other; Jackson *et al.*, 2000). DNA sequencing showed that *avrPphB* is chromosomally located on a 106-kb GEI (PPHGI-1) that shares features with integrative and conjugative elements (ICElands) and PAI in diverse bacteria (Hacker and Carniel, 2001, Van der Meer and Sentchilo, 2003) (Figure 6.2b). The PPHGI-1 island carrying *avrPphB* is lost from 1302A during infection of the resistant bean cv. Tendergreen that expresses the *R3* resistance gene, which recognizes *avrPphB* (Pitman *et al.*, 2005). The island is not lost during infection of the susceptible cv. Canadian Wonder, which lacks *R3*. Excision of PPHGI-1 from chromosomal *att* loci was mediated by an island-encoded *xerC*-like integrase, which was strongly upregulated in both susceptible and resistant plants. This important observation demonstrates the selection pressure imposed by *avrPphB*-based resistance (Figure 6.3). This mechanism of overcoming host resistance is very problematic for agriculture.

6.3.2. *virPphA* PAI

Virulence gene *virPphA* was identified on a plasmid-borne PAI of *P. syringae* pv. *phaseolicola* (Jackson *et al.*, 1999). Strains cured of a 154-kb plasmid lost virulence toward beans causing the HR in previously susceptible cultivars. Restoration of virulence was accomplished by a 30-kb region that contained previously identified *avr* genes and three new putative virulence genes, one of which (*virPphA*) partially restored virulence alone. Sequencing also revealed the presence of insertion sequences IS*100* from *Yersinia* and transposase Tn*501* from *P. aeruginosa*. The proximity of several *avr* and *vir* genes, together with mobile elements and a significantly lower G+C content than the *P. syringae* core genome, signified the region as a PAI on the plasmid.

Homologues of *virPphA* have been detected in other pathovars of *P. savastanoi* and *P. syringae*. Jackson *et al.* (2002) found that in two strains of *P. savastanoi*, the *virPphA* gene is potentially on a PAI, being located next

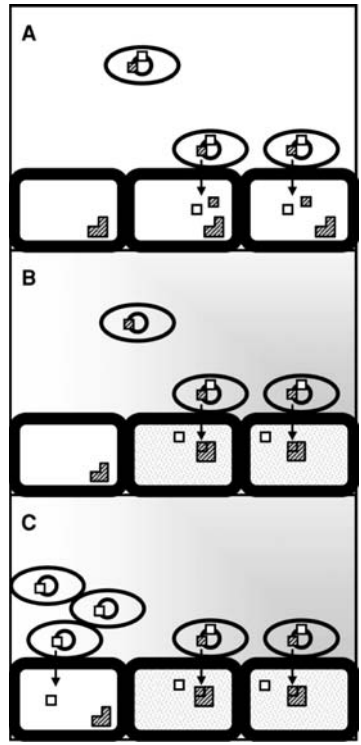


Figure 6.3. The evolution of a virulent pathogen within a resistant host via loss of PPHGI-1. (A) Following infection of bean leaves by bacteria, the pathogen secretes effector proteins (open and hatched squares) into healthy plant cells (white) to try and shut down plant defenses and to release nutrients. (B) The effector protein AvrPphB (hatched square), which is carried on PPHGI-1, is recognized by beans expressing the *R3* resistance gene via the R3 resistance protein (hatched structure). Recognition of AvrPphB activates plant resistance (stippled boxes), leading to the production of an antimicrobial environment outside the plant cell. The antimicrobial environment triggers upregulation of the bacterial PPHGI-1 *xerC* integrase, which is essential for island integration/excision and loss of the island from the bacterial cell. Some newly formed bacterial cells lose PPHGI-1 and consequently *avrPphB*. (C) Cells that have lost PPHGI-1/*avrPphB* deliver other effectors (open squares) that are not recognized by the plants' defense surveillance mechanisms and promote virulence. The plants' metabolism is diverted in favor of bacterial growth and the bacterial cells lacking PPHGI-1 can begin to multiply, eventually to become the dominant genotype.

to another effector *avrPphD* and transposase genes. However, in *P. syringae* pv. *tomato* DC3000 the homologue (*avrPtoB*) is located in a region of the chromosome surrounded by transporter genes that does not appear to be a GEI.

Rivas *et al.* (2005) carried out a comparative analysis of the *virPphA* PAI in a number of *P. syringae* pv. *phaseolicola* strains. They concluded that the PAI is generally well conserved but a 9.5-kb fragment carrying *avrPphF* was shown to be absent from the genome of strain 1448A. The left junction of the deleted region consisted of a chimeric transposable element generated from the fusion of homologues of IS*1492* from *P. putida* and IS*1090* from *Ralstonia eutropha*. The conservation of this border over a number of isolates suggests that the *avrPphF* deletions were mediated by the activity of a chimeric mobile element. Evidence was also presented that most of the strains belonging to races lacking *avrPphF* were derived from a common *P. syringae* pv. *phaseolicola* strain after a unique deletion event in the PAI. This evidence for partial island deletion is presumably due to strong selection pressure to maintain genes such as *virPphA* that promote virulence.

6.3.3. *pop* Islands

The devastating wilt pathogen, *Ralstonia solanacearum* strain GMI1000, encodes three genes, *popP1*, *popP2*, and *popP3*, whose protein products are similar to the type III cysteine protease effectors AvrRxv/YopJ. *popP1* and *popP2* are located on a 79-kb GEI in the *R. solanacearum* chromosome, along with phage-related genes and pathogenicity genes related to hemagglutinins and the *pthG* gene of *Erwinia herbicola* pv. *gypsophilae* (Lavie *et al.*, 2004). *popP3* is on an 83-kb chromosomal GEI that also carries phage-related genes. The *popP* genes appeared to be present mainly in African and Asiaticum isolates, but usually were not found in Americanum isolates.

6.4. OTHER GENOMIC ISLANDS

6.4.1. *Pseudomonas aeruginosa* Pathogenicity Islands (PAPI)

P. aeruginosa causes disease in animals and plants, and a transposon knockout in one gene, RL003, caused a decreased virulence in mouse and plants. Mapping of the gene to the PA14 genome revealed that it is present in two sites; sequence analysis identified these as GEI, designated PAPI-1 and -2. PAPI-1 carries at least 19 virulence factors found in other bacteria (He *et al.*, 2004). Interestingly, the *P. syringae* pv. *phaseolicola* PPHGI-1 island

appears to be structurally similar to PAPI-1 and similar in size. Both islands are located in a $tRNA^{Lys}$ locus next to $tRNA^{Pro}$ and both encode a type IV secretion system and contain integrase genes. PAPI-2 carries the type III effector *exoU*, which is important for pathogenesis. Knockouts of numerous genes on PAPI-1 and RS06 (a hypothetical protein) on PAPI-2 demonstrated a role in virulence to plants.

6.4.2. *Xanthomonas* LPS Islands

Patil and Sonti (2004) identified a 12- to 13-kb region in the chromosome of *Xanthomonas* plant pathogens that exhibit atypical DNA G+C content and with an altered codon usage for ORF. The island contains genes encoding lipopolysaccharide (LPS) synthesis/modification genes and is required for the virulence of *X. oryzae* pv. *oryzae* on host plants. The island does not contain an integrase gene, but does carry transposases. This LPS GEI is present in most but not all *X. oryzae* pv. *oryzae* strains. However, in those strains lacking these LPS genes is an island containing genes similar to *X. axonopodis* pv. *citri* LPS genes or a hybrid between the two different LPS gene clusters.

6.5. PHYTOTOXIN GENOMIC ISLANDS

Phytotoxins are products of plant pathogens that contribute to disease development but are usually not required for basic pathogenicity toward the host plant. Among the best characterized bacterial phytotoxins are those produced by *P. syringae* (Bender *et al.*, 1999). Here we describe some of the evidence suggesting that a number of these toxins are present on GEI.

6.5.1. Coronatine

The principal symptom of coronatine (COR) is a diffuse chlorosis (yellowing) that can be induced on a wide variety of plant species and contributes to the virulence of several *P. syringae* pathovars (Alarcón-Chaidez *et al.*, 1999). In *P. syringae* the gene cluster encoding COR is generally plasmid-encoded and is made up of two distinct components: the polyketide coronafacic acid (CFA) and an ethylcyclopropyl amino acid derived from isoleucine, coronamic acid (CMA). These are separated by a 3.4-kb regulatory region (Alarcón-Chaidez *et al.*, 1999; Figure 6.2c). The COR gene cluster has several features of a GEI: The G+C ratio of CMA and CFA gene regions are lower than the normal

range for *P. syringae,* the COR gene cluster may be chromosomal as observed in pv. *maculicola*, and the cluster is flanked by multiple IS elements.

Recent sequence analysis from Bell *et al.* (2004) has shown the presence of part of a coronatine cluster on a potentially mobile island in *Erwinia carotovora* subsp. *atroseptica* (see genome analysis section).

6.5.2. Syringomycin and Syringopeptin

Syringomycin and syringopeptin are lipodepsipeptide toxins produced by *P. syringae* pv. *syringae* via a non-ribosomal peptide synthetase system, which have been shown to be major virulence determinants (Scholz-Schroeder *et al.*, 2001). Based on reverse transcriptional polymerase chain reaction (PCR) and bioinformatics analysis, Wang *et al.* (2006) demonstrated that the 132-kb *syr-syp* GEI consists of 21 genes divided into five polycistronic operons and eight monocistronic genes. A tripartite resistance-nodulation-cell division (RND) transporter system, called the PseABC efflux system, has been identified at the left border of the *syr-syp* GEI (Kang and Gross, 2005). A mutant with an insertion in the *pseC* gene showed a larger reduction in syringopeptin secretion (67%) than in syringomycin secretion (41%) compared to the parental strain. This PseABC efflux system is the first RND transporter system described for *P. syringae* and has an important role in secretion of syringomycin and syringopeptin.

6.5.3. Tabtoxin

Tabtoxin is produced by various isolates of *P. syringae* (Mitchell, 1991) and is the precursor for the biologically active tabtoxinine-β-lactam. The genes for tabtoxin biosynthesis and resistance are clustered together in a region that is highly conserved among tabtoxin-producing pathovars of *P. syringae* (Kinscherf *et al.*, 1991). This region is physically unstable, and spontaneous deletions occurred that rendered the bacterium both tabtoxin deficient and tabtoxin sensitive in bioassays and eliminated disease symptoms on bean. DNA sequence analysis revealed that the biosynthetic genes of tabtoxin reside adjacent to $tRNA^{LysC}$ gene in *P. syringae* BR2 (Kinscherf and Willis, 2005).

6.5.4. Phaseolotoxin

Phaseolotoxin acts as a virulence factor causing chlorosis by inhibiting the enzymatic activity of ornithine carbamoyl transferase in host plants. The *argK-tox* locus encodes genes involved in phaseolotoxin production that are

clustered together (Zhang *et al.*, 1993). Analysis of the codon usage and the G+C content at the codon third position, in conjunction with phylogenomic analysis, have shown that the *argK-tox* gene cluster expanded its distribution by horizontal transfer into the genomes of two *P. syringae* pathovars (pv. *actinidiae* and pv. *phaseolicola*) from bacterial species distantly related to *P. syringae* (Sawada *et al.*, 2002).

6.5.4. Thaxtomin

A group of the Gram-positive, filamentous Streptomycetes cause the globally important potato scab disease. A PAI has recently been described in *Streptomyces* species (Kers *et al.*, 2005), which appears to be responsible for the emergence of new plant pathogenic *Streptomyces* species in agricultural systems. The PAI was shown to contain the thaxtomin biosynthetic pathway, a putative tomatinase gene, numerous mobile genetic elements, and the *nec1* gene, which encodes a necrogenic protein that is a virulence factor. The authors were able to mobilize the PAI on a 660-kb DNA element from *S. turgidiscabies* to the non-pathogen species *S. coelicolor* and *S. diastatochromogenes*. Acquisition of the PAI conferred a pathogenic phenotype on *S. diastatochromogenes,* but not on *S. coelicolor*.

6.6. GEIslets – SMALL MOBILE REGIONS OR 'LARGE GEI' PROGENITORS/BREAKDOWNS?

Not all GEI may be large genetic elements since there are examples of different single effector genes that seem to be inserted into the same genomic location in different genomes. For example, work by Arnold *et al.* (2001) found several type III effectors, including *avrPpiA*, embedded into conserved regions of DNA on the chromosome or plasmids of some *P. syringae* pv. *pisi* strains. This conservation of flanking sequences has implications for the evolution of pathogenicity, suggesting that specific regions of the bacterial genome act as sites for their integration or excision. Analysis of the DNA flanking *avrPpiA* revealed the gene is located on an interesting area of the genome (Arnold *et al.*, 2000). The gene co-locates with genes similar to bacteriophage and transposase genes on a 4.5-kb region that has inserted within the DNA repair gene *rulB*. These are located within an 8.5-kb fragment in the chromosome that is unique to *P. syringae* pv. *pisi*. These observations suggest multiple rounds of mobility that could be the breakdown or establishment of a GEI. The *avrPpiA*-like effector gene *avrRpm1* from *P. syringae* pv. *maculicola* (Dangl *et al.*, 1992) is also next to a

truncated *rulB* gene (Rohmer *et al.*, 2003). However, *avrRpm1* is located on a 40-kb plasmid pFKN that is important for virulence and which also carries the effector gene *avrPphE* and a plant-inducible gene of unknown function. Another copy of *avrPphE* is present on the chromosome and appears to be part of a transposon. pFKN can integrate and excise from the chromosome at low frequency by homologous recombination of the two *avrPphE* transposons. As the authors suggested, the integration of pFKN into the chromosomal region may be seen as an evolutionary process determining the formation of a new PAI in the chromosome of *P. syringae* pv. *maculicola*. Thus, the observed pFKN chromosomal insertion event seen in pv. *maculicola* appears to be similar to an event that happened in pv. *pisi* much longer ago.

These examples illustrate PAI mosaicism, for example, insertions within insertions. A further example of this phenomenon can be seen by comparison of PPHGI-1 with related regions in the genomes of *P. syringae* pv. *syringae* B728a and pv. *tomato* DC3000 (Pitman *et al.*, 2005). The conserved backbone of the islands is disrupted by variable gene cassettes in each pathovar, containing *avrPphB* in 1302A, a copper and arsenic resistance gene cluster in B728a, and a number of insertions carrying genes predicted to encode candidate type III effector proteins in DC3000. Primers designed to amplify the variable region carrying *avrPphB* in PPHGI-1 generated variable-sized products from 21 strains representing seven pathovars of *P. syringae*. Six amplicon groups were identified; one contained the effector *avrRpt2* from *P. syringae* pv. *tomato* strain JL1065.

6.7. SEQUENCE ANALYSIS TO IDENTIFY ISLANDS

Traditionally, GEI were identified because of a change in an obvious phenotype, such as change of virulence through chromosomal deletion or plasmid loss or due to genomic rearrangements (Blum *et al.*, 1994; Jackson *et al.*, 1999; Pitman *et al.*, 2005). The explosion of genome sequencing data that is now available for plant pathogenic bacteria has led to many more predictions of GEI.

6.7.1. *Xylella* and *Xanthomonas* Comparative Genomics

Pioneering studies have been carried out with the pathogens *Xylella* and *Xanthomonas* to identify GEI and pathogenicity genes. A comparative analysis of the genomes of *Xylella axonopodis* pv. *citri* and *Xanthomonas campestris* pv. *campestris* was done by screening the genome sequences for regions of the genome that have contiguous genes with unusual BLAST

matches (Lima *et al.*, 2005). Forty percent of genes in the two chromosomes have best hit matches to distantly related organisms. Most genes formed clusters that were designated as unusual best-match islands (UBI). Each strain had five exclusive UBI, but shared parts of 30 other UBI. Most UBI contained mobility genes such as transposases or integrases and genes related to virulence factors. Pathogenicity genes included type II and type III protein secretion systems and the xanthan extracellular polysaccharide genes.

Further computational analysis of four bacterial genomes of the Xanthomonadaceae family revealed unique genes that may be involved in adaptation, pathogenicity, and host specificity (Moreira *et al.*, 2005). The genomes analyzed were *Xylella fastidiosa* 9a5c (Xf-9a5c), *Xylella fastidiosa* Temecula 1 (Xf-temecula), *X. axonopodis* pv. *citri*, and *X. campestris* pv. *campestris*. Eighteen unique genes were located inside PinDel-8 (Putative Insertion/Deletion events) of Xf-9a5c and PinDel-4 of *X. axonopodis* pv. *citri*, which were not present in *X. campestris* pv. *campestris* or Xf-temecula. These unique genes comprised a region that is homologous to the SPI-7 PAI from *Salmonella enterica* serovar Typhi. The authors speculated that the presence of SPI-7 in *X. axonopodis* pv. *citri* and XF-9a5c resulted from two separate lateral transfer events, which is consistent with the hypothesis that SPI-7 may be a mobile element (Pickard *et al.*, 2003).

A microarray approach was also used to analyze the genomes of 12 strains of *Xylella fastidiosa*, an important pathogen of many plants including fruit trees and coffee (Nunes *et al.*, 2003). Genes that were upregulated on Periwinkle Wilt medium or *Xylella* defined medium were identified. The genomic location of these genes was identified and found to be clustered to discrete regions of the genome. Many regions exhibited the hallmarks of GEI (Van Sluys *et al.*, 2003), although some were integrated phages or a plasmid. Five GEI (ranging from 9 to 67 kb) and six putative GEI were identified. *Xf* GI1 (37 kb) encodes 41 hypothetical protein genes, a hemolysin-like gene, a lipid-A biosynthesis gene, and a virulence associated protein gene, *vapE*. *X. fastidiosa* GI2 (67 kb) encodes eight transcriptional regulators and harbors genes encoding ketoreductases and dehydrogenases. This island carries an integrase similar to the one carried on the *clc* ICEland. The three smaller GEI in *X. fastidiosa* carry genes encoding hypothetical proteins, although GI5 also carries a copy of *vapE*. The putative GEI encode a number of genes for stress resistance (e.g., copper). A 15.7-kb island was identified in the genome of the Pierce's disease *X. fastidiosa* strain Temecula and carries a hemagglutinin gene (Van Sluys *et al.*, 2003).

6.7.2. *Pseudomonas syringae*

Comparison of the complete genome sequences of *P. syringae* pv. *syringae* B728A and pv. *tomato* DC3000 showed that many of the 976 genes unique to B728A were found in 14 GEI, nine of which appear to be associated with integration into tRNA loci (Feil *et al.*, 2005). GEI structure varied, with one resembling a prophage and another being similar to the copper and arsenic resistance plasmid pKLC102 of *P. aeruginosa* PA01.

Bioinformatics approaches assessing the G+C content of genomes (Gao and Zhang, 2006) has led to the identification of new GEI. Chen (2006) found GEI in *Agrobacterium tumefaciens*, *Ralstonia solanacearum*, *X. axonapodis* pv. *citri*, *X. campestris* pv. *campestris*, *Xylella fastidiosa*, and *P. syringae* pv. *tomato* that had a set of conserved features. For example, *P. syringae* pv. *tomato* contained four GEI (PsyGI1–4) with a G+C content lower than that of the whole genome. PsyGI-1 contains four transposases and a virulence-associated protein. PsyGI-2 contains 86 genes, including seven T3SS effector genes and two integrases situated near the 5′ junction. PsyGI-3 has a $tRNA^{Gln}$ at the 3′ junction, two integrases, four transposases, and three methyl-accepting chemotaxis proteins. PsyGI-4 contains four phage integrases, 10 transposases, and four T3SS effectors.

6.7.3. *Erwinia*

GEI have also been identified from the genome sequence of a plant pathogenic enterobacterium, *E. carotovora* subsp. *atroseptica* (Bell *et al.*, 2004). Eleven putative horizontally acquired GEI (HAI) were determined based on hallmarks of recent gene transfer. Particularly interesting were two islands, HA12 and HA17. HAI7 carries genes that resemble a plasmid conjugation system in *E. coli* and the pathogenicity-related type IV secretion system locus (T4SS) in *A. tumefaciens*. HAI2 showed a high level of conservation of both sequence and gene order to the SPI-7 PAI in *S. enterica* serovar Typhi. However, the SPI-7 pathogenicity determinant *virB* is absent from the *E. carotovora* subsp. *atroseptica* island. In its place are genes highly similar to the *cfa* gene cluster in *P. syringae*, which encode polyketide synthetases of the CFA component of the phytotoxin coronatine. However, genes encoding CMA synthetases of coronatine were not found. Transposon insertions in either island, in the T4SS or in the *cfa* genes, caused a significant reduction in virulence (Bell *et al.*, 2004). This combination of a functional analysis with genome mining is important as it moves beyond simple descriptive analyses.

6.8. DELETION AND MOVEMENT OF ISLANDS

The *P. syringae* pv. *phaseolicola* GEI, PPHGI-1, forms a circular episome after excision from the chromosome (Pitman *et al.*, 2005). An important question we can now ask is whether this episome is capable of moving between bacteria. This has implications in the area of horizontal gene transfer, especially if GEI contain pathogenicity or virulence genes that can be spread between bacterial populations. Plant-microbe biologists can gain valuable insights into this area by reviewing the work done on the mobility of GEI in animal pathogenic bacteria. This will help us develop strategies to test experimentally whether GEI from plant pathogens are mobile. Recent work on the high-pathogenicity island (HPI), which codes for an iron uptake system, has shown that it is present and highly conserved in various Enterobacteriaceae, suggesting its recent acquisition by lateral gene transfer (Lesic and Carniel, 2005). By co-culturing donor and recipient *Yersinia pseudotuberculosis* strains at low temperatures (4°C), the authors were able to observe transfer of the HPI. This transfer was not observed during normal culture conditions for the bacteria (28°C). They also investigated the theory that transfer of a genetic element may be increased when the bacteria are in an environment where expression of the genes carried by the mobile element is required. As HPI encodes the yersiniabactin system, which provides a growth advantage under iron-limiting conditions, the transfer experiments were conducted with the presence of an iron chelator in the culture media, and an eightfold increase in HPI transfer was observed.

Another approach used to show movement of HPI was to 'trap' it in a conjugative plasmid and use this as the mechanism of mobility. Antonenka *et al.* (2005) engineered an RP4 promiscuous conjugative shuttle plasmid to contain an *asn* tRNA gene to create an *attB* site in the plasmid. The HPI island was able to form a cointegrate with this plasmid by site-specific recombination at the *attB* site. This cointegrate was then transferred by conjugation to a new host where the HPI excised from RP4 and integrated into an unoccupied $tRNA_{Asn}$ gene in the genome. As the authors concluded, 'trapping' of GEI and subsequent shuffling to new hosts by conjugative replicon carrying a suitable *att*B site could be applied to other functional integrative elements.

When considering the possibility of the movement of GEI between bacteria, an important question to be addressed is whether expression of genes on the island is affected by the host background and whether this will consequently affect the evolution of islands within different hosts. Wilson and

Nickerson (2006) carried out a study aimed at addressing this question using a 15-kb island from *S. enterica* serovar Typhimurium. The island, STM4305, was cloned into a broad-host range vector and transferred to 10 other Gram-negative hosts belonging to eight different genera of gamma- and alpha-proteobacterial hosts. Gene expression was then studied using a combination of reverse transcriptase PCR and reporter gene promoter fusions. The results showed that the genes were differentially expressed in other hosts in a manner that displays a pattern based on evolutionary relationship: That is, the island genes are not expressed in certain hosts, and these hosts belong to more evolutionarily distant genera of the alpha-proteobacterial group. This type of systematic analysis of island gene expression in a range of plant pathogenic bacteria could lead to significant insights into the mechanisms that regulate gene expression in different hosts and therefore when an island may confer a selective advantage.

6.9. CONCLUSIONS

GEI are being recognized more frequently in plant pathogenic bacteria, either via functional characterization or by in silico analysis of genome sequences. Most islands identified so far are important for pathogenicity, although some are likely to be important for survival in adverse environments (e.g., arsenic resistance in *P. syringae* pv. *syringae* B728a ICEland). While these islands play an important positive role in bacterial fitness in plant hosts, they are also important in the context of crop resistance as demonstrated by the *P. syringae* pv. *phaseolicola* PPHGI-1 ICEland. Greater knowledge is needed of the functions of core genes identified on these islands as well as candidate virulence factors. As He *et al.* (2004) demonstrated in *P. aeruginosa*, core genes do appear to play a role in virulence (e.g., single-stranded DNA-binding protein). But, are they acting directly as virulence factors or are they somehow involved in stability of the island or regulating gene expression? Furthermore, how are these islands disseminated between bacterial plant pathogens? There is currently little knowledge of transfer mechanisms.

There is a clear requirement to identify island-encoded virulence factors that are recognized as avirulence genes in host cultivars. Loss of an island that encodes an avirulence factor can have huge consequences for crop resistance because virulent strains will rapidly emerge within the population if resistance is monogenic. Thus, we need to understand the mechanisms that influence integration and excision of islands into bacterial genomes. This includes the molecular mechanisms in the bacterial cell and any potential

environmental influences, such as plant products that can upregulate excisionases like *xerC*, as observed in PPHGI-1.

Finally, there is also a need to understand how these islands evolve. Several islands appear to have a mosaic structure. Do these islands evolve in quantum leaps, such that single near-instantaneous changes occur in the island genome (Groisman and Ochman, 1996)? This phenomenon has recently been described as the mechanism for evolution of T3SS effector genes (Stavrinides *et al.*, 2006). It is also clear that caution should be exercised when analyzing bacterial diversity via potential island sequences, either because islands can be lost or because their genotypes can be similar but variable. In summary, we now have a firm footing on island identification, but we still have much to do in exploration of the interior.

REFERENCES

Alarcón-Chaidez, F. J., Peñaloza-Vázquez, A., Ullrich, M., and Bender, C. L. (1999). Characterization of plasmids encoding the phytotoxin coronatine in *Pseudomonas syringae*. *Plasmid*, **42**, 210–20.

Alfano, J. R., and Collmer A. (1997). The type III (Hrp) secretion pathway of plant pathogenic bacteria: trafficking harpins, Avr proteins, and death. *J Bacteriol*, **179**, 5655–62.

Alfano, J. R., Charkowski, A. O., Deng, W., *et al.* (2000). The *Pseudomonas syringae* Hrp pathogenicity island has a tripartite mosaic structure composed of a cluster of type III secretion genes bounded by exchangeable effector and conserved effector loci that contribute to parasitic fitness and pathogenicity in plants. *Proc Natl Acad Sci USA*, **97**, 4856–61.

Antonenka U., Nolting C., Heesemann J., and Rakin A. (2005). Horizontal transfer of *Yersinia* high-pathogenicity island by the conjugative RP4 *attB* target-presenting shuttle plasmid. *Mol Microbiol* **57**, 727–34.

Araki, H., Tian, D., Goss, E. M., *et al.* (2006). Presence/absence polymorphism for alternative pathogenicity islands in *Pseudomonas viridiflava*, a pathogen of *Arabidopsis*. *Proc Natl Acad Sci USA*, **103**, 5887–92.

Arnold, D. L., Jackson, R. W., and Vivian, A. (2000). Evidence for the mobility of an avirulence gene, *avrPpiA1*, between the chromosome and plasmids of races of *Pseudomonas syringae* pv. *pisi*. *Mol Plant Pathol*, **1**, 195–9.

Arnold, D. L., Jackson, R. W., Fillingham, A. J., *et al.* (2001). Highly conserved (DNA) sequences flank avirulence genes: isolation of novel avirulence genes from *Pseudomonas syringae* pv *pisi*. *Microbiology*, **147**, 1171–82.

Bell, K. S., Avrova, A. O., Holeva, M. C., *et al.* (2002). Sample sequencing of a selected region of the genome of *Erwinia carotovora* subsp. *atroseptica* reveals

candidate phytopathogenicity genes and allows comparison with *Escherichia coli*. *Microbiology*, **148**, 1367–78.

Bell, K. S., Sebaihia, M., Pritchard, L., *et al.* (2004). Genome sequence of the enterobacterial phytopathogen *Erwinia carotovora* subsp. *atroseptica* and characterization of virulence factors. *Proc Natl Acad Sci USA*, **101**, 11105–10.

Bender, C. L., Alarcón-Chaidez, F., and Gross, D. C. (1999). *Pseudomonas syringae* phytotoxins: mode of action, regulation, and biosynthesis by peptide and polyketide synthetases. *Microbiol Mol Biol Rev* **63**, 266–92.

Blum, G., Ott, M., Lischewski, A., *et al.* (1994). Excision of large DNA regions termed pathogenicity islands from tRNA-specific loci in the chromosome of an *Escherichia coli* wild-type. *Infect Immun*, **62**, 606–14.

Bogdanove, A., Kim, J. F., Wei, Z., *et al.* (1998). Homology and functional similarity of an *hrp*-linked pathogenicity locus, *dspEF*, of *Erwinia amylovora* and the avirulence locus *avrE* or *Pseudomonas syringae* pathovar *tomato*. *Proc Natl Acad Sci USA*, **95**, 1325–30.

Büttner, D., and Bonas, U. (2002). Getting across – bacterial type III effector proteins on their way to the plant cell. *EMBO J*, **21**, 13–22.

Casper-Lindley, C., Dahlbeck, D., Clark, Eszter, T., and Staskawicz, B. J. (2002). Direct biochemical evidence for type III secretion-dependent translocation of the AvrBs2 effector protein into plant cells. *Proc Natl Acad Sci USA*, **99**, 8336–41.

Charity, J. C., Pak, K., Delwiche, C. F., and Hutcheson, S. W. (2003). Novel exchangeable effector loci associated with the *Pseudomonas syringae hrp* pathogenicity island: evidence for integron-like assembly from transposed gene cassettes. *Mol Plant-Microbe Interact*, **16**, 495–507.

Chen, L. L. (2006). Identification of genomic islands in six plant pathogens. *Gene*, **374**, 134–41.

Dangl, J. L., Ritter, C., Gibbon, M. J., *et al.* (1992). Functional homologs of the *Arabidopsis* RPM1 disease resistance gene in bean and pea. *Plant Cell*, **4**, 1359–69.

Deng, W. L., Rehm, A. H., Charkowski, A. O., Rojas, C. M., and Collmer, A. (2003). *Pseudomonas syringae* exchangeable effector loci: Sequence diversity in representative pathovars and virulence function in *P. syringae* pv. *syringae* B728a. *J Bacteriol*, **185**, 2592–602

Feil, H., Feil, W. S., Chain, P., *et al.* (2005). Comparison of the complete genome sequences of *Pseudomonas syringae* pv. *syringae* B728a and pv. *tomato* DC3000. *Proc Natl Acad Sci USA*, **102**, 11064–9.

Frederick, R. D., Ahmad, M., Majerczak, D. R., *et al.* (2001). Genetic organization of the *Pantoea stewartii* subsp. *Stewartii* hrp gene cluster and sequence analysis of the *hrpA*, *hrpC*, *hrpN* and *wtsE* operons. *Mol Plant-Microbe Interact*, **14**, 1213–22.

Gao, F., and Zhang, C.-T. (2006). GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences. *Nucleic Acids Res*, **34** http://nar.oxfordjournals.org/cgi/content/full/34/suppl_2/W686doi: 10.1093/nar/gkl040.

Genin, S., and Boucher, C. (2004). Lessons learned from the genome analysis of *Ralstonia solanacearum*. *Annu Rev Phytopathol*, **42**, 107–34.

Groisman, E. A., and Ochman, H. (1996) Pathogenicity islands: bacterial evolution in quantum leaps. *Cell*, **87**, 791–4.

Guo, M., Manulis, S., Mor, H., and Barash, I. (2002). The presence of diverse IS elements and an *avrPphD* homologue that acts as a virulence factor on the pathogenicity plasmid of *Erwinia herbicola* pv. *gypsophilae*. *Mol Plant-Microbe Interact*, **15**, 709–16.

Hacker, J., and Carniel, E. (2001). Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep*, **2**, 376–81.

Hacker, J., and Kaper, J.B. (2000). Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol*, **54**, 641–79.

Hacker, J., Bender, L., Ott, M., *et al.* (1990). Deletions of chromosomal regions coding for fimbriae and hemolysins occur *in vitro* and *in vivo* in various extraintestinal *Escherichia coli* isolates. *Microb Pathog* **8**, 213–25.

He, J., Baldini, R. L., Deziel, E., *et al.* (2004). The broad host range pathogen *Pseudomonas aeruginosa* strain PA14 carries two pathogenicity islands harboring plant and animal virulence genes. *Proc Natl Acad Sci USA*, **101**, 2530–5.

Hentschel, U., and Hacker, J. (2001). Pathogenicity islands: the tip of the iceberg. *Microbes Infect*, **3**, 545–8.

Jackson, R. W., Athanassopoulos, E., Tsiamis, G., *et al.* (1999). Identification of a pathogenicity island, which contains genes for virulence and avirulence, on a large native plasmid in the bean pathogen *Pseudomonas syringae* pathovar *phaseolicola*. *Proc Natl Acad Sci USA*, **96**, 10875–80.

Jackson, R. W., Mansfield, J. W., Arnold, D. L., *et al.* (2000). Excision from tRNA genes of a large chromosomal region, carrying *avrPphB*, associated with race change in the bean pathogen, *Pseudomonas syringae* pv. *phaseolicola*. *Mol Microbiol*, **38**, 186–97.

Jackson, R. W., Mansfield, J. W., Ammouneh, H., *et al.* (2002). Location and activity of members of a family of *virPphA* homologues in pathovars of *Pseudomonas syringae* and *P. savastanoi*. *Mol Plant Pathol*, **3**, 205–16.

Kang, H., and Gross, D. C. (2005). Characterization of a resistance-nodulation-cell division transporter system associated with the *syr-syp* genomic island of *Pseudomonas syringae* pv. *syringae*. *Appl Environ Microbiol*, **71**, 5056–65.

Keen, N. T. (1990). Gene-for-gene complementarity in plant–pathogen interactions. *Ann Rev Genet*, **24**, 421–40.

Kers, J. A., Cameron, K. D., Joshi, M. V., *et al.* (2005). A large, mobile pathogenicity island confers plant pathogenicity on *Streptomyces* species. *Mol Microbiol*, **55**, 1025–33.

Kim, J. F., Charkowski, A. O., Alfano, J. A., Collmer, A., and Beer, S. V. (1998). Sequences related to transposable elements and bacteriophages flank avirulence genes of *Pseudomonas syringae*. *Mol Plant-Microbe Interact*, **11**, 1247–52.

Kim, J. G., Park , B. K., Yoo, C. H., *et al.* (2003). Characterization of the *Xanthomonas axonopodis* pv. *glycines* Hrp pathogenicity island. *J Bacteriol*, **185**, 3155–66.

Kinscherf, T. G., and Willis, D. K. (2005). The biosynthetic gene cluster for the beta-lactam antibiotic tabtoxin in *Pseudomonas syringae*. *J Antibiot (Tokyo)*, **58**, 817–21.

Kinscherf, T. G., Coleman, R. H., Barta, T. M., and Willis, D. K. (1991). Cloning and expression of the tabtoxin biosynthetic region from *Pseudomonas syringae*. *J Bacteriol*, **173**, 4124–32.

Lavie, M., Seunes, B., Prior, P., and Boucher, C. (2004). Distribution and sequence analysis of a family of type III-dependent effectors correlate with the phylogeny of *Ralstonia solanacearum* strains. *Mol Plant-Microbe Interact*, **17**, 931–40.

Lesic, B., and Carniel, E. (2005). Horizontal transfer of the high-pathogenicity island of *Yersinia pseudotuberculosis*. *J Bacteriol*, **187**, 3352–8.

Lima, W. C., Van Sluys, M., and Menck, C. F. M. (2005). Non-Gamma-Proteobacteria gene islands contribute to the *Xanthomonas* genome. *OMICS*, **9**, 160–72.

Mitchell, R. E. (1991). Implications of toxins in the ecology and evolution of plant pathogenic microorganisms: bacteria. *Experientia*, **47**, 791–803.

Mor, H., Manulis, S., Zuck, M., *et al.* (2001). Genetic organization of the *hrp* gene cluster and *dspAE/BF* operon in *Erwinia herbicola* pv. *gypsophilae*. *Mol Plant-Microbe Interact* **14**, 431–6.

Moreira, L. M., De Souza, R. F., Digiampietri, L. A., Da Silva, A. C., and Setubal, J. C. (2005). Comparative analyses of *Xanthomonas* and *Xylella* complete genomes. *OMICS*, **9**, 43–76

Noël, L., Thieme, F., Nennstiel, D., and Bonas, U. (2002). Two novel type III-secreted proteins of *Xanthomonas campestris* pv. *vesicatoria* are encoded within the *hrp* pathogenicity island. *J Bacteriol*, **184**, 1340–8.

Nunes, L. R., Rosato, Y. B., Muto, N. H., *et al.* (2003). Microarray analyses of *Xylella fastidiosa* provide evidence of coordinated transcription control of laterally transferred elements. *Genome Res*, **13**, 570–8.

Oh, C-S., Kim, J. F., and Beer, S. V. (2005). The Hrp pathogenicity island of *Erwinia amylovora* and identification of three novel genes required for systemic infection. *Mol Plant Pathol*, **6**, 125–38.

Patil, P. B., and Sonti, R. V. (2004). Variation suggestive of horizontal gene transfer at a lipopolysaccharide (*lps*) biosynthetic locus in *Xanthomonas oryzae* pv. *oryzae*, the bacterial leaf blight pathogen of rice. *BMC Microbiol*, **4**, 40.

Pickard, D., Wain, J., Baker, S., *et al.* (2003). Composition, acquisition, and distribution of the Vi exopolysaccharide-encoding *Salmonella enterica* pathogenicity island SPI-7. *J Bacteriol*, **185**, 5055–65.

Pitman, A. R., Jackson, R. W., Mansfield, J. W., *et al.* (2005). Exposure to host resistance mechanisms drives evolution of bacterial virulence in plants. *Curr Biol*, **15**, 2230–5.

Rivas, L. A., Mansfield, J., Tsiamis, G., Jackson, R. W., and Murillo, J. (2005). Changes in race-specific virulence in *Pseudomonas syringae* pv. *phaseolicola* are associated with a chimeric transposable element and rare deletion events in a plasmid borne pathogenicity island. *Appl Environ Microbiol*, **71**, 3778–85.

Rohmer, L., Kjemtrup, S., Marchesini, P., and Dangl, J. L. (2003). Nucleotide sequence, functional characterization and evolution of pFKN, a virulence plasmid in *Pseudomonas syringae* pathovar *maculicola*. *Mol Microbiol*, **47**, 1545–62.

Rossier, O., Van den Ackerveken, G., and Bonas, U. (2000). HrpB2 and HrpF from *Xanthomonas* are type III-secreted proteins and essential for pathogenicity and recognition by the host plant. *Mol Microbiol*, **38**, 828–38.

Sawada, H., Kanaya, S., Tsuda, M., *et al.* (2002). A phylogenomic study of the OCTase genes in *Pseudomonas syringae* pathovars: the horizontal transfer of the *argK-tox* cluster and the evolutionary history OCTase genes on their genomes. *J Mol Evol*, **54**, 437–57.

Scholz-Schroeder, B. K., Hutchison, M. L., Grgurina, I., and Gross, D. C. (2001). The contribution of syringopeptin and syringomycin to virulence of *Pseudomonas syringae* pv. *syringae* strain B301D on the basis of *sypA* and *syrB1* biosynthesis mutant analysis. *Mol Plant-Microbe Interact*, **14**, 336–48.

Stavrinides, J., Ma, W., and Guttman, D. (2006). Terminal reassortment drives the quantum evolution of type III effectors in bacterial pathogens. *PLoS Pathog*, **2**, e104.

Sugio, A., Yang, B., and White, F. F. (2005). Characterization of the *hrpF* pathogenicity peninsula of *Xanthomonas oryzae* pv. *oryzae*. *Mol Plant-Microbe Interact*, **18**, 546–54.

Toth, I. K., Bell, K. S., Holeva, M. C., and Birch, P. R. J. (2003). Soft rot erwiniae: From genes to genomes. *Mol Plant Pathol*, **4**, 17–30.

Van der Meer, J. R., and Sentchilo, V. (2003). Genomic islands and the evolution of catabolic pathways in bacteria. *Curr Opin Biotechnol*, **14**, 248–54.

van Dijk, K., Fouts, D. E., Rehm, A. H., *et al*. (1999). The Avr (effector) proteins HrmA (HopPsyA) and AvrPto are secreted in culture from *Pseudomonas syringae* pathovars via the Hrp (type III) protein secretion system in a temperature- and pH-sensitive manner. *J Bacteriol*, **181**, 4790–7.

Van Sluys, M. A., de Oliveira, M. C., Monteiro-Vitorello, C. B., *et al*. (2003). Comparative analyses of the complete genome sequences of Pierce's disease and citrus variegated chlorosis strains of *Xylella fastidiosa*. *J Bacteriol*, **185**, 1018–26.

Wang, N., Lu, S. E., Yang, Q., Sze, S. H., and Gross, D. C. (2006). Identification of the *syr-syp* box in the promoter regions of genes dedicated to syringomycin and syringopeptin production by *Pseudomonas syringae* pv. *syringae* B301D. *J Bacteriol*, **188**, 160–8.

Wilson, J. W., and Nickerson, C. A. (2006). A new experimental approach for studying bacterial genomic island evolution identifies island genes with bacterial host-specific expression patterns. *BMC Evol Biol*, **6**, 2.

Zhang, Y., Rowley, K. B., and Patil, S. S. (1993). Genetic organization of a cluster of genes involved in the production of phaseolotoxin, a toxin produced by *Pseudomonas syringae* pv. *phaseolicola*. *J Bacteriol*, **175**, 6451–8.

CHAPTER 7

Prophage Contribution to *Salmonella* Virulence and Diversity

Sébastien Lemire, Nara Figueroa-Bossi, and Lionello Bossi

7.1. INTRODUCTION

Salmonellae are enteric pathogenic bacteria that infect a vast spectrum of animal species from reptiles to mammals. The genus is highly diversified, comprising more than 2,500 serovars. An even greater diversity exists at the strain level as a result of countless combinations of genomic differences in individual isolates. Strain-specific variations are likely to influence aspects of the organism biology, such as the adaptation to specific hosts or environments, the tropism for certain organs or tissues, and the degree of pathogenicity. The emergence in recent years of epidemic clones that have become predominant, displacing pre-existing strains, confirms that strain diversification is an ongoing process. A leading mechanism promoting diversity in *Salmonella* genomes is lysogenization by temperate phages. Most strains harbor multiple prophages in variable numbers and combinations. Prophages modify the properties of the host bacterium in various ways, from expressing functions that directly influence pathogenicity, to improving the bacterium's ability to outgrow competitors or to resist killing by superinfecting phages. Unlike toxin-producing phages of other bacterial species, most *Salmonella* prophages contribute to virulence through the synergistic action of multiple factors playing subtle, often redundant roles. The scope of this chapter is to review the various facets of phage-mediated modification of *Salmonella*, as well as recent advances in the characterization of phage-borne virulence determinants.

7.2. *SALMONELLA* DIVERSITY

Current classification divides the genus *Salmonella* into two species, *Salmonella bongori* and *Salmonella enterica*. The latter is subdivided into seven subspecies designated by roman numerals (Beltran *et al.*, 1988; Boyd *et al.*, 2003; Crosa *et al.*, 1973; Reeves *et al.*, 1989). More than 99% of isolates from humans and other warm-blooded hosts belong to *Salmonella enterica* subspecies I (Kingsley and Bäumler, 2000). Further classification by the Kauffman-White scheme is based on a serological differentiation of surface antigens variability and sorts bacteria into serovars (Beltran *et al.*, 1988; Le Minor, 1984). In its latest update, 2451 *Salmonella* serovars are listed, of which about 60% belong to *S. enterica* subspecies I (also called subspecies *enterica*; Popoff *et al.*, 2004). Further diversity within any given serovar was evident and has led to development of a variety of subtyping schemes. A standard procedure classifies isolates on the basis of their sensitivity to a reference collection of bacteriophages. This technique has allowed to classify *Salmonella enterica* serovars of medical and veterinarian relevance such as serovars Typhimurium and Enteritidis into 207 and 27 different phage types, respectively (Anderson *et al.*, 1977; Ward *et al.*, 1987).

7.3. *SALMONELLA* VIRULENCE

Although sometimes carried asymptomatically, Salmonellae are most often associated with acute illness in humans and warm-blooded animals. Infection typically results from the consumption of contaminated food or water. Bacteria overcome the gastric barrier, colonize the ileum, and penetrate the ileal epithelium through the M cells of the Peyer's patches. The infection can then either remain localized to the intestinal tract or spread systemically. The localized infection causes enterocolitis, a syndrome characterized by widespread mucosal damage and massive fluid release into the gut lumen (Santos *et al.*, 2002; Zhang *et al.*, 2003). Although some bacteria enter the blood stream, bacteremia is generally transient during enterocolitis except in young or immunocompromised hosts. In the systemic infection, the intestinal epithelium suffers little damage as bacteria disseminate to lymph nodes, liver, and spleen, where they multiply. The increase in bacterial load leads to a severe septicemia that can be fatal if untreated (Fierer and Guiney, 2001; Ohl and Miller, 2001; Zhang *et al.*, 2003). Occasionally, the infection recedes and can evolve into an asymptomatic carrier state that constitutes a relevant source of *Salmonella* circulation in the food chain (Fierer and Guiney, 2001). The course of the infection is determined by interplay of

bacterial and host factors. *Salmonella* serovars Typhimurium and Enteritidis, two serovars with a broad host range, cause gastroenteritis in most of their hosts (e.g., humans and cattle) and a systemic disease resembling typhoid fever in rodents (Bäumler *et al.*, 1998; Kingsley and Bäumler, 2000; Rabsch *et al.*, 2002; Uzzau *et al.*, 2000). In contrast, *Salmonella* serovars adapted or restricted to specific hosts typically produce a systemic infection. *Salmonella* serovar Typhi, which is strictly pathogenic to humans and is the causative agent of typhoid fever, is the most representative member of this class (Pascopella *et al.*, 1995).

A complex regulatory network assures the optimal spatial and temporal expression of virulence genes during *Salmonella* infection. At the heart of the system are two-component regulators that link transcriptional control to environmental cues, often acting as part of a signal transduction cascade. One of these systems is the PhoP-PhoQ master regulator, which couples expression of a large number of genes to intracellular growth (Altier, 2005; Groisman, 2001; Groisman and Mouslim, 2006). Phosphorylated PhoP exerts its regulatory activity either directly by binding the promoter regions of target genes, or indirectly by activating expression of regulatory proteins. A large fraction of *Salmonella* virulence genes lie in 'genomic islands' (GEI), which are thought to originate by horizontal transmission. This pathogenic arsenal includes two separate type III secretion systems (T3SS) encoded by pathogenicity islands 1 (SPI-1) and 2 (SPI-2). These are specialized, needle-like organelles that mediate the translocation of effector proteins from inside the bacterium directly into the host cell cytosol (Aizawa, 2001). Some of the effectors delivered by the SPI-1–T3SS induce cytoskeleton rearrangements that promote bacterial uptake (Collazo and Galan, 1997; Galan and Curtiss, 1989). This system is therefore essential for *Salmonella* to be able to invade the intestinal epithelium. SPI-1 is present throughout the genus *Salmonella* and its acquisition is thought to have been a major event leading to the divergence from the genus *Escherichia* (Bäumler, 1997; Groisman and Ochman, 1996; Hensel, 2004).

Salmonella intracellular replication takes place within a vacuolar compartment that forms upon bacterial entry. In spite of the harshness of this environment, Salmonellae have evolved the ability for intraphagosomal survival and multiplication. This property, which is critical for systemic dissemination, relies on the activity of a second T3SS encoded on SPI-2 (Ochman *et al.*, 1996; Shea *et al.*, 1996). While the function of many SPI-2-delivered effectors remains to be determined, some of them have been associated with cellular vesicle trafficking. In particular, some of these proteins were proposed to inhibit the phagosome-lysosome fusion and the trafficking of

NADPH-containing vesicles, thus protecting bacteria from the oxidative burst (Vazquez-Torres and Fang, 2001). SPI-2 is absolutely required for *Salmonella* to cause a systemic infection. Its acquisition by an ancestral strain marked the separation between *S. bongori*, which lacks SPI-2, and *S. enterica*. Additional genes involved in intra-macrophage survival or participating in the enteric phase of the infection are found on other pathogenicity islands (SPI-3, SPI-4, and SPI-5) and various islets around the chromosome (Blanc-Potard *et al.*, 1999; Wong *et al.*, 1998; Wood *et al.*, 1998). The recurrent association of virulence determinants with horizontally acquired DNA has intimately linked the study of *Salmonella* virulence to that of genome evolution.

7.4. POLY-LYSOGENY IN *SALMONELLA*

Most, if not all, *Salmonella* isolates carry a variable number of prophages integrated in their genomes. These prophages can carry genes that influence the properties of the host strain including pathogenic functions. Unlike the vast majority of acquired DNA elements, the association of viral DNA with the host chromosome is reversible, since the phage can excise, replicate, and spread to other strains. Given the propensity of phage DNA to engage in recombinational exchanges, phage trafficking has the potential for constituting a primary source of variation in *Salmonella* genomes. This topic was subject of recent comprehensive reviews (Bossi and Figueroa-Bossi, 2005; Brüssow *et al.*, 2004; Ehrbar and Hardt, 2005; Waldor and Friedman, 2005). This chapter mainly focuses on the most novel aspects.

At this point a brief look at some basic facts of the lysogenic process will be useful. Lysogenization constitutes one of the two developmental pathways of temperate phages. Rather than multiplying, the phage enters a dormant state, which most often involves integrating its DNA into the host chromosome. In the resulting lysogen, a phage-encoded repressor shuts off transcription of all genes required for viral development. The repressor can also act in trans on newly injected phage DNA, rendering the strain immune against superinfection by the phage it bears. Although not the primary choice of a temperate phage (Kourilsky, 1973), lysogenization is the mandatory outcome of the encounter of a phage with a bacterial population. As the number of bacterial cells becoming infected by newly replicated viruses increases, a lysogen will inevitably appear in the population. Being resistant to superinfection, the latter will prosper and eventually replace the entire bacterial pool. Thus, phage killing acts as a selection regimen that leads to the accumulation of lysogens (Bossi *et al.*, 2003). The association of prophage with host DNA can come to an end under conditions that activate the host's RecA protein.

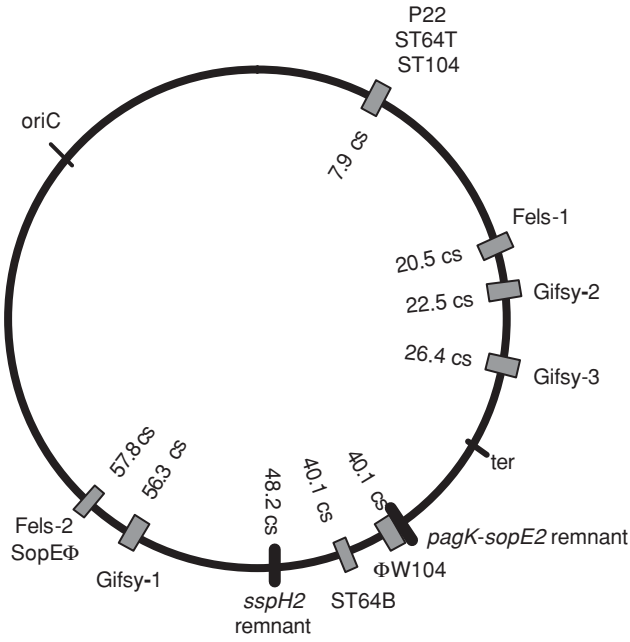


Figure 7.1. Schematic diagram showing the positions of prophages and prophage remnants in the *Salmonella enterica* serovar Typhimurium chromosome.

Activated RecA mediates the relief of repression, causing the prophage DNA to become excised and enter the lytic cycle. This process, generally referred to as prophage induction, is stimulated by various treatments, but it also occurs spontaneously at low, but nevertheless significant, frequencies. There can be little doubt that prophage induction is a major source of continuous phage release in the environment (Boyd, 1950). The phenomenon might be particularly relevant in bacteria causing diarrheal diseases, where it could reach considerable proportions following treatments with antibiotics that target DNA gyrase and elicit RecA activation (Matsushiro *et al.*, 1999; Zhang *et al.*, 2000).

Prophage insertion occurs via a site-specific recombination event involving a circular form of phage DNA and the chromosome (Campbell, 1993). As a result, the core segment of the recombining sequences is duplicated at the two ends of the prophage. The positions of eight prophages, or prophage families, found in *Salmonella* serovar Typhimurium are shown in Figure 7.1. No one strain has yet been found that carries all of the eight prophages; rather, most strains harbor smaller subsets (typically four or five) in a variable

assortment. This variability reflects differences in prophage distribution. Some prophages are found in most strains (Gifsy-1 and Gifsy-2), others have a more limited distribution (SopEΦ, St64B, P22), and others still are specific to certain isolates (Fels-1, Gifsy-3). Extensive sequence variability also exists within each of the prophages as a result of widespread genetic shuffling (see later discussion for some specific examples), which renders naming these elements a particularly challenging exercise. We have found no better solution than to associate the name of the phage with its chromosomal attachment site and to specify the strain that phage is from as a superscript. Thus, for example, Gifsy-1^{LT2} and Gifsy-1^{DT104} designate phages inserted in the 5′ region of the *lepA* gene in *Salmonella* serovar Typhimurium strains LT2 and DT104, respectively. As we will describe later, despite the common root, these two phages differ significantly from each other.

7.4.1. The Gifsy Prophages

Gifsy-1 and Gifsy-2 are lambdoid prophages present in the vast majority of *Salmonella enterica* serovar Typhimurium strains. The two elements were initially identified genetically during a study of *recB* suppressors in laboratory strain LT2 (Figueroa-Bossi *et al.*, 1997) and subsequently their genomes were sequenced as part of the strain's genomic sequence project (McClelland *et al.*, 2001). Work with a different strain, ATCC14028, identified a third prophage that appeared genetically related to Gifsy-1 and Gifsy-2 and was named Gifsy-3 (Figueroa-Bossi *et al.*, 2001). All threes prophages contain *recE* gene orthologs that are normally repressed but that can be activated by mutations resulting in the suppression of *recB* recombination defects (Figueroa-Bossi *et al.*, 1997; Lemire, 2006). Although structurally and morphologically related to bacteriophage λ, these three prophages undergo induction by a different mechanism. Unlike λ, where induction results from RecA-stimulated autocatalytic cleavage of the *cI* repressor (Crowl *et al.*, 1981; Roberts and Roberts, 1975), the Gifsy phage repressors are not cleaved; rather, they are inactivated by the binding of small antirepressor proteins. The genes encoding these proteins are located outside the immunity regions within a small operon under the control of the SOS response master regulator, LexA (Lemire, 2006).

Of the three phages, Gifsy-2 is the most widely distributed, being found in a number of different serovars (Boyd *et al.*, 2003; Porwollik *et al.*, 2004; Reen *et al.*, 2005; Thomson *et al.*, 2004), albeit sometimes in a deleted or extensively rearranged version (Bacciu *et al.*, 2004; Figueroa-Bossi *et al.*, 2006). For example, the element found at the Gifsy-2 attachment site in serovar Typhi shares

with its Typhimurium counterpart only the integrase gene and the immunity modules (Parkhill *et al.*, 2001; Deng *et al.*, 2003; Thomson *et al.*, 2004). An approximately 1.5-kb insert whose last 200 bp are 74% identical to the right end of Gifsy-2^{LT2} can be detected at the Gifsy-2 *att* site in *S. bongori*, suggesting that an ancestor of the phage already existed at the time of the separation of the two *Salmonella* species (http://www.sanger.ac.uk/Projects/Salmonella/). Although subjected to variability within the species, Gifsy-2 is nonetheless highly conserved within *S. enterica* serovar Typhimurium (Reen *et al.*, 2005; http://www.sanger.ac.uk/Projects/Salmonella/). In contrast, prophage Gifsy-1 is restricted to a limited number of serovars and shows extensive strain-to-strain variability within *S. enterica* serovar Typhimurium (Chiu *et al.*, 2005; Reen *et al.*, 2005). Intriguingly, such variability affects mostly the prophage portion involved in lysogenic regulation and immunity (Bossi and Figueroa-Bossi, 2005; Bossi and Figueroa-Bossi, unpublished). Thus, for example, Gifsy-1 phages from model strains LT2, ATCC14028, and SL1344 all have different immunity modules. In strain LT2, the left third of the genome of Gifsy-2, including the immunity determinants, is duplicated at the corresponding position of Gifsy-1, presumably as a result to of an intrachromosomal recombination/conversion event (Lemire, 2006; McClelland *et al.*, 2001). Thus, Gifsy-1^{LT2} has the same immunity as Gifsy-2 (the latter being conserved throughout the serovar). Furthermore, Gifsy-1^{SL1344} has the immunity of Gifsy-3^{ATCC14028} (Bossi and Figueroa-Bossi, 2005). Clearly, the variability in the immunity region offers the basis for further gene shuffling, since a strain lysogenic for Gifsy-1 will be still infectable by Gifsy-1 phage from another strain. Recombination will allow reassortment of sequences and generation of new strains and phage.

Prophage Gifsy-3 shows a highly limited distribution. A PCR-based survey of the occupancy of Gifsy-3 attachment site detected an insert in two of the 21 serovar Typhimurium strains from the SARA collection (Beltran *et al.*, 1991) and in none of a group of 72 clinical strains isolated in France in 2002 (Figueroa-Bossi, Weill, Grimont and Bossi, unpublished).

7.4.2. Fels-2 and SopEΦ

Fels-2 and SopEΦ are related members of the P2 phage family found in serovar Typhimurium strains LT2 and SL1344, respectively (Hardt *et al.*, 1998; McClelland *et al.*, 2001). The two phages are inserted at the same position at the 3′ end of the *ssrA* gene encoding tmRNA (Pelludat *et al.*, 2003). Their main difference is the presence of an inversion system in the tail fiber region of Fels-2, which is replaced by a DNA segment containing the *sopE*

virulence gene in SopEΦ. Interestingly, the Fels-2 invertase was recently reported to participate in *Salmonella* flagellar phase variation (Kutsukake *et al.*, 2006). These two phages have a scattered distribution in the *S. enterica* serovar Typhimurium, where SopEΦ has been associated with a group of strains responsible for epidemic outbreaks in the United Kingdom and former East Germany in the 1970s and 1980s (Mirold *et al.*, 1999; Prager *et al.*, 2000). Their presence in other *Salmonella* serovars has been assessed to a limited extent (Thomson *et al.*, 2004). Fels-2 and SopEΦ are heteroimmune. Infection of strain LT2 by SopEΦ can have different outcomes including the triggering of the induction of the resident Fels-2 prophage, resulting in the replacement of Fels-2 by SopEΦ, or leading to the formation of tandem lysogens (Bossi and Figueroa-Bossi, 2005).

7.4.3. ΦW104

Isolates from a variety of *Salmonella* serovars, including Typhi, Paratyphi, Gallinarum, and Enteritidis, carry a lambdoid prophage at 40.1 centisomes (Deng *et al.*, 2003; Figueroa-Bossi *et al.*, 2006; Parkhill *et al.*, 2001; http://www.sanger.ac.uk/Projects/Salmonella/). In serovars Typhimurium, this prophage is specifically associated with the DT104 lineage (Hermans *et al.*, 2005, 2006). This prophage is inducible and the resulting virus capable to infect and lysogenize naïve strains (Bossi and Figueroa-Bossi, 2005). We refer to it as ΦW104, a designation that will be also used to indicate some of its relatives in other serovars. The ΦW104 prophage is inserted immediately adjacent to a prophage remnant carrying the *pagK* and *pagO* loci (see later discussion).

7.4.4. St64B

First identified in an epidemic serovar Typhimurium DT64 strain (Mmolawa *et al.*, 2003a), prophage ST64B is present in a large fraction of serovar Typhimurium strains. The prophage is integrated at the *serU* locus. In some strains, including the original DT64 host and laboratory strains SL1344 and ATCC14029, ST64B is in a defective form because of a frame-shift mutation in a gene involved in tail assembly (Figueroa-Bossi and Bossi, 2004). The mutation does not prevent the prophage from undergoing induction. However, viral morphogenesis cannot be completed, resulting in the accumulation of tailless virions (Mmolawa *et al.*, 2003a). Because of spontaneous reversion of the frame-shift mutation, some rare infective particles occur

in the phage population. These are rapidly amplified upon co-culturing the lysogenic strain with a strain susceptible to ST64B infection (Figueroa-Bossi and Bossi, 2004). Spontaneous induction of the ST64B prophage was found to be significantly stimulated in *Salmonella* strains defective in Dam methylase (Alonso *et al.*, 2005; Balbontin *et al.*, 2006). It is currently unclear whether this phenotype reflects a direct role of methylation on the lysogenic control mechanism or it is an indirect consequence of chronic low-level RecA activation caused by Dam deficiency (Alonso *et al.*, 2005).

7.4.5. P22, ST64T, ST104

Bacteriophage P22 occupies a relevant place in the history of microbial genetics both as a model system (Susskind and Botstein, 1978; Zinder and Lederberg, 1952). Among the phages described here, P22 is the only one capable of generalized transduction. The phage also provides a classical example for lysogenic conversion as it modifies the antigenic formula of the host bacterium upon establishing itself as a prophage. Serotype conversion results from the addition of a glucosyl residue to galactose moieties in the O-antigen repeats of LPS, which is mediated by the products of the *gtrABC* operon located at the right end of the prophage map (Fukazawa and Hartman, 1964; Vander Byl and Kropinski, 2000). The change prevents further binding of the phage to its O-antigen receptor, thus contributing to the exclusion of superinfecting phage from the same group. Serotype conversion may also influence the interaction of lysogenic bacteria with the animal host as shown for other phage. However, to date, this possibility lacks experimental support. The P22 genome includes a sequence identical to the last 46 bp of the *thrW* gene for $tRNA_2^{Thr}$, which carries the recombination site used for integration into the host chromosome (Leong *et al.*, 1985).

Several lines of evidence point to the high incidence of P22-related prophages in *Salmonella* genomes (Schicklmaier *et al.*, 1998, 1999; Schicklmaier and Schmieger, 1995). The DNA sequences of two such phages, ST64T and ST104, released from a definite type strains DT64 and DT104, respectively, were published recently (Mmolawa *et al.*, 2003b; Tanaka *et al.*, 2004). A study examining the configuration downstream from the *thrW* gene in *S.* Typhimurium isolates from the SARA collection found a P22-like insert in seven of 21 strains analyzed. The incidence of a P22-like insert was higher in isolates of epidemiological relevance. In particular, all strains from definite type DT104 and DT120 tested were found to contain a P22-related insert (Figueroa-Bossi *et al.*, 2007).

7.4.6. Fels-1

Prophage Fels-1 is located between the loci *ybjP* and STM0930 in the chromosome of serovar Typhimurium strain LT2 (McClelland *et al.*, 2001). A PCR-based survey of the occupancy of Fels-1 *attB* site in 72 clinical isolates from serovar Typhimurium and in 21 strains from the SARA collection (Beltran *et al.*, 1991) showed the site to be free of inserts in all but the LT2 entry of the collection (Figueroa-Bossi *et al.*, 2007). An earlier study surveying the distribution of Fels-1-associated *nanH* gene in a larger sample of serovar Typhimurium strains found only strain LT2 to contain the gene. In contrast *nanH* was present in about 60% of isolates from *Salmonella enterica* subspecies III (formerly *S. arizonae*) (Hoyer *et al.*, 1992). Assuming the tight association of *nanH* with the Fels-1 prophage, these results indicate the rarity of Fels-1 in serovar Typhimurium and suggest that strain LT2 acquired the prophage from a subspecies III strain.

7.4.7. Prophage Remnants

Three separate regions in the *Salmonella* serovar Typhimurium genome show sequence similarity to phage tail operons, suggesting that they represent the remains of ancestral prophages. Two of these islets, found in various versions throughout the *Salmonella* complex, contain genes that can be linked to pathogenicity. One region of approximately 17 kb located at 40.1 cs harbors the *pagK*, *mig-3*, *pagO*. and *sopE2* loci (Amavisit *et al.*, 2003; Bakshi *et al.*, 2000; Gunn *et al.*, 1998; Mirold *et al.*, 2001; Stender *et al.*, 2000). The other segment of about 11 kb at 48.2 cs harbors the *sspH2* gene (Miao *et al.*, 1999, 2003; Miao and Miller, 2000). Some features of the *pagK-sopE2* locus, namely the presence of two related *int* genes separated by approximately 10 kb of DNA and ending near the same 23-bp sequence, suggest that it actually comprises two tandem elements that were incorporated in the course of separate events.

7.5. PROPHAGE-ENCODED VIRULENCE GENES

Some of the prophages described earlier encode one or more factors that can be linked to pathogenicity by any of various criteria, including direct evidence from animal infection studies, type III secretion and host-specific regulation (Bacciu *et al.*, 2004; Figueroa-Bossi and Bossi, 1999; Figueroa-Bossi *et al.*, 2001; Hardt *et al.*, 1998; Ho *et al.*, 2002; Saitoh *et al.*, 2005). These loci and their relevant properties are summarized in Table 7.1.

Table 7.1. *Virulence genes on Salmonella phages*

Gene name	Phage	*Salmonella* serovar	Function and/or features	Distribution	References or web source
sodCI	Gifsy-2 ΦW104	Typhimurium Choleraesuis Enteritidis Gallinarum	Periplasmic CuZnSOD Protects bacteria against reactive oxygen species	Absent in sv. Typhi and Paratyphi A sequenced genomes	Chiu *et al.*, 2005; De Groote *et al.*, 1997; Deng *et al.*, 2003; Fang *et al.*, 1999; Figueroa-Bossi *et al.*, 2006; Figueroa-Bossi and Bossi, 1999; http://www.salmonella.org/genomics/sequences.html; http://www.sanger.ac.uk/Projects/Salmonella/; McClelland *et al.*, 2004; McClelland *et al.*, 2001; Parkhill *et al.*, 2001
sopE	*sopE*Φ Gifsy-2 ΦW104	Typhimurium Hadar Enteritidis Gallinarum	Type III secreted protein; G-nucleotide exchange factor; stimulates host cell invasion	Scattered distribution within sv. Typhimurium absent in sv. Choleraesuis sequenced genome	Chiu *et al.*, 2005; Deng *et al.*, 2003; Hardt *et al.*, 1998; http://www.sanger.ac.uk/Projects/Salmonella/; McClelland *et al.*, 2004; McClelland *et al.*, 2001; Mirold *et al.*, 2001; Mirold *et al.*, 1999; Parkhill *et al.*, 2001; Thomson *et al.*, 2004

(*continued*)

Table 7.1 *(continued)*

Gene name	Phage	*Salmonella* serovar	Function and/or features	Distribution	References or web source
gtgE	Gifsy-2 Gifsy-2 remnant	Typhimurium Choleraesuis Enteritidis	Contributes to virulence in mice; function unknown	Absent in sv. Typhi, Paratyphi, Gallinarum sequenced genomes	Chiu *et al.*, 2005; Deng *et al.*, 2003; Figueroa-Bossi *et al.*, 2006; Ho *et al.*, 2002; http://www.sanger.ac.uk/Projects/Salmonella/; McClelland *et al.*, 2004; McClelland *et al.*, 2001; Parkhill *et al.*, 2001
sseI (*srfH*; *gtgB*)	Gifsy-2	Typhimurium Choleraesuis Abortusovis	Type III secreted protein; binds actin cross-linking protein filamin; binds TRIP6 stimulating macrophage motility	Absent in all other sequenced *Salmonella* genomes	Bacciu *et al.*, 2004; Chiu *et al.*, 2005; Deng *et al.*, 2003; McClelland *et al.*, 2004; McClelland *et al.*, 2001; Miao *et al.*, 2003; Parkhill *et al.*, 2001; Worley *et al.*, 2006
gogB	Gifsy-1	Typhimurium Choleraesuis	Type III secreted LRR protein; localizes in the host cell cytoplasm; function unknown	Absent in all other sequenced *Salmonella* genomes	Chiu *et al.*, 2005; Coombes *et al.*, 2005; Deng *et al.*, 2003; McClelland *et al.*, 2004; McClelland *et al.*, 2001; Parkhill *et al.*, 2001; Uzzau *et al.*, 2001
gipA	Gifsy-1	Typhimurium Choleraesuis	Involved in mouse Peyer's patch colonization	Absent in all other sequenced *Salmonella* genomes	Chiu *et al.*, 2005; Deng *et al.*, 2003; http://www.sanger.ac.uk/Projects/Salmonella/; McClelland *et al.*, 2004; McClelland *et al.*, 2001; Parkhill *et al.*, 2001; Stanley *et al.*, 2000

sspH1	Gifsy-3	Typhimurium	Type III-secreted LRR protein; localizes in the host cell nucleus; downregulates NF-κB-dependent transcription	Highly infrequent	Figueroa-Bossi *et al.*, 2001; Haraga and Miller, 2003; Haraga and Miller, 2006; Miao *et al.*, 1999
artA/B	Gifsy-1 Prophage remnant	Typhimurium Typhi Paratyphi A	Pertussis-like toxin; ADP-ribosyltransferase toxin	Sporadic in sv. Typhimurium; absent in all other sequenced *Salmonella* genomes	Chiu *et al.*, 2005; Deng *et al.*, 2003; http://www.sanger.ac.uk/Projects/Salmonella/; McClelland *et al.*, 2004; McClelland *et al.*, 2001; Parkhill *et al.*, 2001; Saitoh *et al.*, 2005
astA	Gifsy-2AO	Abortusovis	Heat-stable enterotoxin	Highly sporadic	Bacciu *et al.*, 2004
gogA	Gifsy-1	Typhimurium	Similar to *pipA*; function unknown		Figueroa-Bossi *et al.*, 2001
gtgA	Gifsy-2	Typhimurium Choleraesuis	Similar to *pipA*; function unknown		Figueroa-Bossi *et al.*, 2001
	ΦW104	Gallinarum Enteritidis			
gogD	Gifsy-1	Typhimurium	*phoP-phoQ*-activated; highly similar to *pagK/J*; function unknown		Figueroa-Bossi *et al.*, 2001; Gunn *et al.*, 1998

(*continued*)

Table 7.1 *(continued)*

Gene name	Phage	*Salmonella* serovar	Function and/or features	Distribution	References or web source
pagJ	Gifsy-3	Typhimurium	*phoP-phoQ*-activated; highly similar to *pagK*; function unknown		Figueroa-Bossi *et al.*, 2001; Gunn *et al.*, 1998; Thomson *et al.*, 2004
sopE2	Prophage remnant	Most serovars	Type III-secreted protein; G-nucleotide exchange factor; stimulates host cell invasion		Bakshi *et al.*, 2000; Friebel *et al.*, 2001; Mirold *et al.*, 2001; Stender *et al.*, 2000
sspH2	Prophage remnant	Most serovars	Type III-secreted LRR protein; binds actin cross-linking protein filamin; binds profilin		Miao *et al.*, 2003; Miao and Miller, 2000; Miao *et al.*, 1999

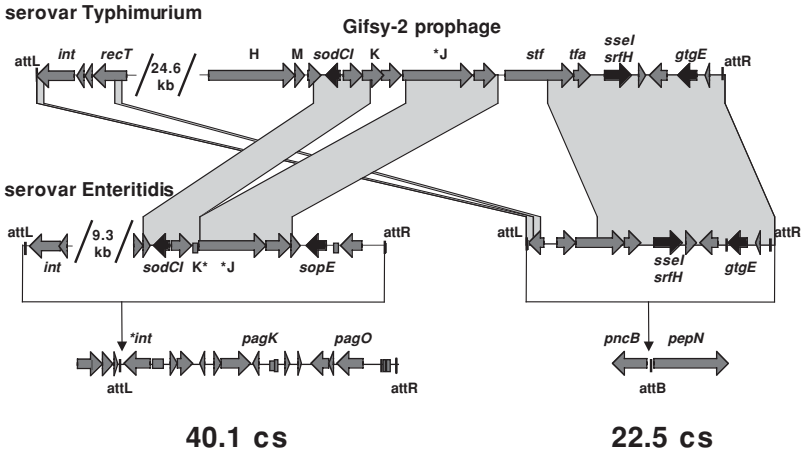


Figure 7.2. Organization of the chromosomal regions containing the *sodCI* gene in *Salmonella enterica* serovars Typhimurium and Enteritidis. The diagram is drawn from data in McClelland *et al.* (2001), from the *Salmonella enterica* serovar Enteritidis strain LK5 sequence project, University of Illinois (http://www.salmonella.org/genomics/sequences.html), and from data produced by the *Salmonella* spp. Comparative Sequencing group at the Sanger Institute (http://www.sanger.ac.uk/Projects/Salmonella/).

7.5.1. Factors Contributing to Virulence in Murine Models

7.5.1.1. SodCI

Possibly the best known virulence factor associated with phage in *Salmonella* is SodCI, a Cu-Zn co-factored periplasmic superoxide dismutase (SOD). The *sodCI* gene is located within the tail-encoding region of the Gifsy-2 prophage in serovars Typhimurium and Choleraesuis and in the corresponding region of a ΦW104-related prophage in serovars Enteritidis and Gallinarum (Figure 7.2). In both serovars Typhimurium and Enteritidis, inactivation of the *sodCI* gene causes an approximate five- to 10-fold attenuation of virulence in a mouse model (Figueroa-Bossi *et al.*, 2006; Uzzau *et al.*, 2002). SodCI is thought to scavenge superoxide anions produced during the macrophage oxidative burst, preventing formation of highly toxic compounds such as hydroxyl radical and peroxynitrite (De Groote *et al.*, 1997). Thus, the SodCI role in virulence is to protect invading bacteria against the host immune response. The protein is synthesized at high levels during *Salmonella* intracellular proliferation (Figueroa-Bossi *et al.*, 2006; Uzzau *et al.*, 2002). Upregulation of a *sodCI::lacZ* fusion in vivo was

recently shown to depend on the PhoP/PhoQ two-component system, suggesting that the *sodCI* promoter is activated by PhoP (Golubeva and Slauch, 2006).

Salmonella contains a second periplasmic [Cu, Zn] SOD, SodCII, encoded by a conserved chromosomal gene under the control of alternative transcription factor σ^S (Fang *et al.*, 1999; Golubeva and Slauch, 2006). SodCII does not contribute to virulence to any significant extent. In part, this reflects the poor expression of the *sodCII* gene in vivo (Figueroa-Bossi *et al.*, 2006; Uzzau *et al.*, 2002). However, SodCII fails to complement the virulence defect of a *sodCI* mutant even when expressed from the *sodCI* promoter, suggesting that SodCII lacks enzymatic features that render SodCI particularly effective at removing macrophage-derived superoxide. The two proteins are 62% identical at the amino acid level and differ in a number of aspects including subunit structure, protease sensitivity, and tightness of association with the periplasm (Gabbianelli *et al.*, 2004; Krishnakumar *et al.*, 2004). A recent study correlated the superior qualities of SodCI in protecting against the oxidative burst with the enzyme's peculiar resistance to protease treatment (Krishnakumar *et al.*, 2007). Conceivably, protease resistance is critically important for allowing the enzyme to remain active in the harsh environment of the phagosome.

An additional *sodC* gene, *sodCIII*, is carried by the Fels-1 phage. This gene is expressed in vitro at levels comparable to *sodC1* (Uzzau *et al.*, 2001). However, like the SodCII enzyme, SodCIII cannot substitute for SodC1 as far as contribution to mouse virulence is concerned (Figueroa-Bossi *et al.*, 2001). It seems possible that the incorporation of *sodC* genes into phage genomes was initially selected for reasons relating to the biology of phage and that *sodCI* subsequently evolved to confer an advantage on the lysogenic host. For example, one might envision that CuZnSOD activity delays prophage induction by scavenging toxic compounds that might trigger prophage induction. The presence of putative *sodC* genes in two prophages from an enterohemorrhagic *E. coli* strain (Perna *et al.*, 2001) would be consistent with this idea.

7.5.1.2. GtgE

The initial characterization of prophage Gifsy-2 contribution to virulence showed that a strain cured for Gifsy-2 is significantly more attenuated than a simple *sodCI* deletion mutant (Figueroa-Bossi and Bossi, 1999; Ho *et al.*, 2002). Subsequent search for additional virulence genes in the prophage genome identified *gtgE*, a gene located at the far right end of the prophage map

(Figure 7.2). Deletion of *gtgE* caused an approximate seven-fold attenuation in competition assays, and the defect could be suppressed by the reintroduction of the locus in single copy (Ho *et al.*, 2002). The *gtgE* gene encodes a 26 kDa protein synthesized constitutively at significant levels by *Salmonella* bacteria growing in vitro or inside host cells (Uzzau *et al.*, 2001). The protein shows no significant similarity to any protein in databases, and to date no clues have been found concerning its function. The individual inactivation of *sodC1* or *gtgE* attenuates virulence less than 10-fold; however, in combination the two mutations act synergistically, causing a nearly 100-fold attenuation, basically resembling the virulence defect of the Gifsy-2-cured strain. This has led to the conclusion that the *sodC1* and *gtgE* genes account for most, if not all, of the contribution of the Gifsy-2 prophage to mouse virulence (Ho *et al.*, 2002).

7.5.2. Phage-encoded Type III-secreted Factors

7.5.2.1. SopE

The *sopE* gene encodes a 25-kDa protein delivered into the epithelial cells by the SPI-1-encoded T3SS (Wood *et al.*, 1996; Hardt *et al.*, 1998). It appears that this gene has been extensively exchanged between prophages in the course of *Salmonella* evolution. It is found in the SopEΦ prophage in serovar Typhimurium, in prophage Gifsy-2 in serovar Hadar (Pelludat *et al.*, 2003), and in ΦW104-related prophage in serovars Gallinarum and Enteritidis (Figueroa-Bossi *et al.*, 2006). In serovar Typhi, the entire SopEΦ prophage appears to have translocated from the *ssrA* locus into a large pathogenicity island (SPI-7) that also includes operons involved in the synthesis of the Vi antigen and type IV pili (Parkhill *et al.*, 2001; Bueno *et al.*, 2004). The prophage contains one intact copy of the 47 bp *att* core sequence at its right end but only the innermost 8-bp portion of the sequence on its left end, suggesting that it has lost the capacity to excise (Parkhill *et al.*, 2001; Pelludat *et al.*, 2003).

SopE is a guanine nucleotide exchange factor that activates at least two members of the RhoGTPase family, Cdc42 and Rac-1 (Hardt *et al.*, 1998). Acting in concert with other SPI-1 secreted proteins, which include SopB, SopE2, and SipA, SopE promotes cytoskeleton rearrangements that favor bacterial penetration, disrupt tight junction structure, and elicit inflammatory responses in various animal models (Boyle *et al.*, 2006; Hapfelmeier *et al.*, 2004; Patel and Galan, 2006; Wallis and Galyov, 2000; Zhang *et al.*, 2002). The functional redundancy of SopE with some of these proteins, particularly

SopE2 with which it shares approximately 70% identity, makes the individual contributions of SopE and SopE2 to virulence rather subtle and might explain the limited distribution of the gene in *Salmonella* serovar Typhimurium. In one study, however, introduction of the *sopE* gene into strain ATCC14028, which normally lacks the locus, was shown to cause a small but significant increase of fluid accumulation in infected bovine ligated ileal loops (Zhang *et al.*, 2002).

7.5.2.2. GogB

The *gogB* gene is located at the right end of Gifsy-1, a region of the prophage that is conserved in all serovar Typhimurium strains that have been surveyed. The gene encodes a 56-kDa leucine-rich repeat (LRR) protein that is substrate for both T3SS encoded in *Salmonella* pathogenicity islands SPI-1 and SPI-2 (Coombes *et al.*, 2005). The GogB protein is translocated into the infected host cell (a process specifically mediated by SPI-2) where it localizes in the cytoplasm. Expression of the *gogB* gene is under the control of the SPI-2-encoded SsrB activator protein. However, the gene is also efficiently expressed in enteropathogenic *E. coli* under conditions that activate LEE pathogenicity island gene expression and type III secretion. Thus, *gogB* constitutes an autonomous module with a versatile adaptability to the regulatory circuitry and secretory apparatus of a wide range of enteric pathogens (Coombes *et al.*, 2005). The target and function of the protein in the host cell remain unknown, and experiments aimed at assessing its contribution of GogB to virulence in mice were inconclusive (Bossi and Figueroa-Bossi, unpublished). However, GogB might participate in a pathway not adequately tested in the mouse model. Significantly, inactivation of *espK*, a *gogB* ortholog present in enterohemorrhagic *E. coli*, was reported to affect bacterial persistence in the intestine of orally infected calves (Vlisidou *et al.*, 2006). The *espK* gene is located within a defective prophage at the corresponding position occupied by *gogB* in Gifsy-1. It directs the synthesis of a protein that is similarly translocated into the host cell cytoplasm by the LEE-encoded T3SS (Vlisidou *et al.*, 2006).

7.5.2.3. SseI/SrfH

A Gifsy-2 gene divergently named *sseI*, *srfH*, and *gtgB* was independently identified by two laboratories as a gene specifying a type III-secreted protein under the control of the SPI-2-encoded SsrB activator protein (Uzzau *et al.*, 2001; Worley *et al.*, 2000). The gene is located near the right end of the prophage and strongly activated in bacteria that proliferate within host

cells or in mouse tissues (Uzzau *et al.*, 2001; Worley *et al.*, 2000). The SseI (SrfH) protein was shown to localize to the polymerizing actin cytoskeleton of eukaryotic cells where it specifically interacts with actin cross-linking protein filamin through its N-terminal domain (Miao *et al.*, 2003). A more recent study unveiled a different type of interaction, involving the host factor TRIP6, a protein that localizes to the plasma membrane and regulates cell adhesion and motility. Based on these findings and on additional evidence, it was proposed that SrfH (SseI) stimulates the phagocyte-mediated systemic spread of bacteria by modulating TRIP6 activity (Worley *et al.*, 2006). The impact of this model is somewhat difficult to reconcile with the absence of the *srfH* (*sseI*) gene in serovars causing systemic infection (e.g., Typhi) and with the lack of a virulence defect in a *srfH* (*sseI*) mutant injected in the mouse (Ho *et al.*, 2002).

7.5.2.4. SspH1

The *sspH1* gene, located at the right end of prophage Gifsy-3 encodes a leucine-rich repeat (LRR) protein translocated into the host cell by both, SPI-1 and SPI-2 T3SS (Miao *et al.*, 1999; Mirold *et al.*, 2001; Zhang *et al.*, 2002). Among all phage-encoded T3SS substrates known to date in *Salmonella*, SspH1 is the only protein that is targeted to the host cell nucleus. SspH1 was shown to suppress production of pro-inflammatory cytokines by inhibiting nuclear factor (NF)-κB-dependent transcription (Haraga and Miller, 2003). This inhibitory activity is exerted through the interaction between the LRR domain of SspH1 and a host factor recently identified as a serine/threonine protein kinase (PKN1) (Haraga and Miller, 2006). Thus, by interacting with PKN1, SspH may inhibit the host inflammatory response and enhance pathogenicity. However, a deletion of the *sspH1* gene did not affect *Salmonella* virulence in a series of assays except in a strain also lacking *sspH2*, which encodes another T3SS-translocated protein of the LRR family (see later discussion). Unlike single mutant strain in *sspH1* or *sspH2*, the *sspH1*/*sspH2* double mutant strain was significantly attenuated in its capacity to elicit enterocolitis in calves (Miao *et al.*, 1999). A study on the occurrence of the *sspH1* gene across the *Salmonella* genus detected the sequence in only a few isolates from *Salmonella enterica* subspecies I; in contrast, the locus was present at a considerably higher frequency in strains from other subspecies (Tsolis *et al.*, 1999).

7.5.2.5. SspH2 and SopE2

Two further T3SS effector proteins are encoded by prophage remnants. SspH2 is an LRR protein of 780 amino acids with an N-terminal domain (first

100 amino acids) nearly identical to that of SseI (SrfH) (Miao and Miller, 2000; Miao *et al.*, 1999). Like the latter, SspH2 co-localizes with the polymerizing actin cytoskeleton interacting with cross-linking protein filamin. In addition, SspH2 interacts with actin-binding protein profilin through its carboxy-terminal domain and inhibits actin polymerization in vitro (Miao *et al.*, 2003). However, the in vivo function of SspH2 remains unknown. The effector protein SopE2 is approximately 70% identical to SopE and also acts as a G-nucleotide exchange factor (Bakshi *et al.*, 2000; Stender *et al.*, 2000). However, the activities of the two proteins are not completely redundant. Unlike SopE, which activates both host signaling molecules Cdc42 and RacA1, SopE2 efficiently activates Cdc42 but not Rac1 (Friebel *et al.*, 2001).

7.5.3. Other Virulence Genes Associated with Gifsy-1

7.5.3.1. A Pertussis Toxin-like Gene in Gifsy-1 Phage from Epidemic DT104

In the past two decades, the majority of human and animal infections by *Salmonella* serovar Typhimurium in industrialized countries have been associated with the multi-drug-resistant DT104 clone (Glynn *et al.*, 1998; Threlfall *et al.*, 1994, 1996). Although antibiotic resistance is likely to have contributed to the spread of the strain, it seems possible that the emergence and high incidence of the clone might partially result from the acquisition of enhanced virulence traits (Carlson *et al.*, 2000, 2001; Glynn *et al.*, 1998). Recently, Saitoh and coworkers (2005) identified the genes for a presumptive ADP-ribosyltransferase toxin, *artA/artB*, in the phage fraction of mitomycin C-treated DT104 cultures. *artA* and *artB* encode proteins with significant sequence similarity to putative pertussis toxin-like subunits of *Salmonella* serovars Typhi and Paratyphi A. Southern hybridization analysis of 12 DT104 isolates using an *artA* probe detected the gene in all 12 strains, whereas only one of 14 non-DT104 strains hybridized to the probe (Saitoh *et al.*, 2005). The authors did not identify the prophage involved. However, the current accessibility to DT104 complete genome sequence at the Sanger Institute allowed us to determine that the *artAB* locus is in fact located within the variable region of prophage Gifsy-1 (Figure 7.3). The locus is positioned to the right of the DNA replication module, downstream from an ORF encoding a putative Q-like anti-terminator protein. This organization is strikingly similar of that of Shiga toxin-encoding phages (H19B and 933W) from enteropathogenic *E. coli*. In H19B and 933W prophages, the Shiga toxin genes are located immediately downstream from the Q gene and maximally expressed during

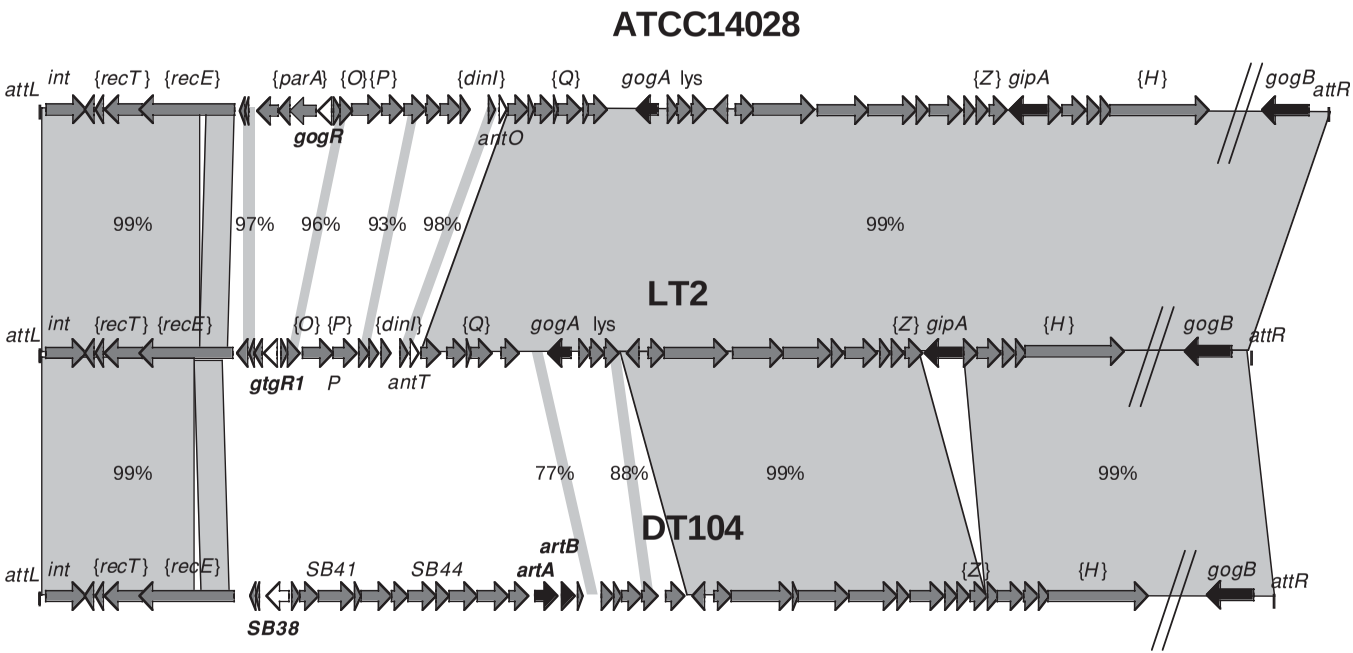


Figure 7.3. Comparison of the Gifsy-1 prophage genomes from *Salmonella enterica* serovar Typhimurium strains LT2 (McClelland *et al.*, 2001), ATCC14028 (Bossi and Figueroa-Bossi, unpublished), and DT104. The DT104 sequence data were produced by the *Salmonella* spp. comparative sequencing group at the Sanger Institute and can be obtained from http://www.sanger.ac.uk/Projects/Salmonella/. Gene designations in curly brackets are based on similarity to known genes in other phages or bacteria. Open arrows indicate genes involved in lysogenic regulation; filled arrows indicate known or putative virulence determinants.

prophage induction (Plunkett *et al.*, 1999; Tyler *et al.*, 2004; Wagner *et al.*, 2001, 2002). A similar coupling of *artAB* expression to lytic development in Gifsy-1^{DT104} would represent the first such example in *Salmonella* where all phage encoded virulence factors known to date are expressed in the lysogenic state. Like the toxin-converting *E. coli* prophages, Gifsy-1^{DT104} exhibits a high rate of spontaneous induction (Bossi *et al.*, 2003). Clearly, although the role of *artA*/*artB* in DT104 virulence remains to be demonstrated, these findings suggest that acquisition of the genes *via* Gifsy-1 phage constituted a relevant step in the emergence of the clone.

7.5.3.2. GipA

Using in vivo expression technology (IVET), J. Slauch and coworkers identified a Gifsy−1^{ATCC14028} locus transiently induced during *Salmonella* colonization of the small intestine (Stanley *et al.*, 2000).This locus, named *gipA*, is needed for proliferation or survival of bacteria in Peyer's patches. A *gipA* null mutant was slightly attenuated in mice when delivered by the oral route but showed no virulence defects when inoculated intraperitoneal. More recently, a *gipA* insertion mutant was found impaired for growth in macrophages (Klumpp and Fuchs, 2007). To date, the function of GipA remains unknown. The protein shows similarity to a family of DNA transposases; however, transposition appeared not to be required for GipA function in Peyer's patches (Stanley *et al.*, 2000). The authors of this work suggested the possibility that GipA is a site-specific recombinase that regulates the expression of virulence genes. We notice that the Gifsy-1^{DT104} prophage lacks *gipA* (Figure 7.3) and that no other copy of this gene can be found elsewhere in the DT104 genome. This might indicate that *gipA* relevance in pathogenicity is restricted to specific hosts or environments.

7.5.3.3. GogA

The *gogA* gene encodes a putative 27-kDa protein showing 72% identity with the PipA protein encoded by pathogenicity island SPI-5 (Wood *et al.*, 1998). A nearly identical copy of *gogA*, designated *gtgA*, is present in the genome of Gifsy-2. Both genes are located in the prophages' late operons in the opposite orientation relative to late operon transcription. This suggests that *gogA* and *gtgA* expression occurs mainly in the lysogenic state. Inactivation of either gene did not significantly affect the ability of *Salmonella* to cause a systemic infection in mice (Figueroa-Bossi *et al.*, 2001). The function of the two proteins is currently unknown.

7.6. GENES INVOLVED IN VIRAL DEVELOPMENT: A CAUTIONARY NOTE

In addition to the genes discussed in the previous sections, some lines of evidence suggest that functions normally devoted to phage replication and morphogenesis might also participate in *Salmonella*-host interactions. Genetic screens for insertions mutants defective in intracellular replication in vitro (Klumpp and Fuchs, 2007), or counter-selected during host infection (Ku *et al.*, 2005; Lawley *et al.*, 2006) have yielded insertions in genes specifying viral functions with no obvious connection to pathogenicity. These include phage DNA replication and recombination genes, as well as genes encoding tail-fiber proteins or proteins required for tail assembly. These data should be interpreted with caution. On one hand, the recurrence of mutations in tail-fiber assembly proteins might suggest that these proteins perform some additional functions that influence growth in the eukaryotic host. The occasional finding of these genes being expressed in vivo would be consistent with this idea (Valdivia and Falkow, 1997). On the other hand, the involvement of other genes connected with the viral lifestyle is more difficult to reconcile. In this respect, we should point out that in our genetic analyses of *Salmonella* prophages we often encountered situations where changes in the phage genome had deleterious effects on bacterial growth. The use of transposons for random mutagenesis has the potential for perturbing the fragile stability of the lysogenic conditions. The consequences from low-level activation of viral functions might be insignificant and pass unnoticed during growth in the laboratory, but they could seriously jeopardize the chances of *Salmonella* to survive and persist in the host environment.

7.7. CONCLUSIONS

One can expect the number and the diversity of prophages within a bacterial genome to be a direct reflection of that organism's lifestyle. Bacterial pathogens occupying secluded niches within their hosts will have reduced opportunity to come in contact with phage. Even if occasionally lysogenized, highly adapted bacteria might receive little benefit from DNA acquisition and will be under selective pressure for loosing the prophage. Thus, for example, obligate intracellular pathogens such as *Mycoplasma* or *Chlamydia* species or stomach-adapted *Helicobacter pylori* completely lack prophages in their genomes (Brüssow *et al.*, 2004). In contrast, enteric bacteria alternating between different hosts and environments are more exposed to phage encounters and can profit from lysogenization for diversifying and

conquering new niches. This is the case for *Salmonella* species whose infection spectrum includes hosts as distinct as reptiles, birds, and mammals. As reviewed here, several *Salmonella* phages harbor virulence loci that 'plug in' the regulatory circuitry of the host bacterium upon lysogenization. Acquisition of these modules can therefore lead to changes that take effect immediately and can confer new properties to the lysogenic cell in a single step. In other cells, phage multiplication and recombination with resident prophages can generate phage variants with new combinations of virulence genes. Overall, there can be little doubt that the interplay of phage diversification and lysogeny constitutes a powerful force driving *Salmonella* evolution and the emergence of new strains.

REFERENCES

Aizawa, S. I. (2001). Bacterial flagella and type III secretion systems. *FEMS Microbiol Lett*, **202**, 157–64.

Alonso, A., Pucciarelli, M. G., Figueroa-Bossi, N., and Garcia-Del Portillo, F. (2005). Increased excision of the *Salmonella* prophage ST64B caused by a deficiency in Dam methylase. *J Bacteriol*, **187**, 7901–11.

Altier, C. (2005). Genetic and environmental control of *Salmonella* invasion. *J Microbiol*, **43**, 85–92.

Amavisit, P., Lightfoot, D., Browning, G. F., and Markham, P. F. (2003). Variation between pathogenic serovars within *Salmonella* pathogenicity islands. *J Bacteriol*, **185**, 3624–35.

Anderson, E. S., Ward, L. R., Saxe, M. J., and De Sa, J. D. (1977). Bacteriophage-typing designations of *Salmonella typhimurium*. *J Hyg (Lond)*, **78**, 297–300.

Bacciu, D., Falchi, G., Spazziani, A., *et al.* (2004). Transposition of the heat-stable toxin astA gene into a gifsy-2-related prophage of *Salmonella enterica* serovar Abortusovis. *J Bacteriol*, **186**, 4568–74.

Bakshi, C. S., Singh, V. P., Wood, M. W., *et al.* (2000). Identification of SopE2, a *Salmonella* secreted protein which is highly homologous to SopE and involved in bacterial invasion of epithelial cells. *J Bacteriol*, **182**, 2341–4.

Balbontin, R., Rowley, G., Pucciarelli, M. G., *et al.* (2006). DNA adenine methylation regulates virulence gene expression in *Salmonella enterica* serovar Typhimurium. *J Bacteriol*, **188**, 8160–8.

Bäumler, A. J. (1997). The record of horizontal gene transfer in *Salmonella*. *Trends Microbiol*, **5**, 318–22.

Bäumler, A. J., Tsolis, R., Ficht, T. A., and Adams, L. G. (1998). Evolution of host adaptation in *Salmonella enterica*. *Infect Immun*, **66**, 4579–87.

Beltran, P., Musser, J. M., Helmuth, R., *et al.* (1988). Toward a population genetic analysis of *Salmonella*: genetic diversity and relationships among strains of serotypes *S. choleraesuis*, *S. derby*, *S. dublin*, *S. enteritidis*, *S. heidelberg*, *S. infantis*, *S. newport*, and *S. typhimurium*. *Proc Natl Acad Sci USA*, **85**, 7753–7.

Beltran, P., Plock, S. A., Smith, N. H., *et al.* (1991). Reference collection of strains of the *Salmonella typhimurium* complex from natural populations. *J Gen Microbiol*, **137**, 601–6.

Blanc-Potard, A. B., Solomon, F., Kayser, J., and Groisman, E. A. (1999). The SPI-3 pathogenicity island of *Salmonella enterica*. *J Bacteriol*, **181**, 998–1004.

Bossi, L., and Figueroa-Bossi, N. (2005). Prophage arsenal of *Salmonella enterica* serovar Typhimurium. In Waldor, M., Friedman, D., and Adhya, S. (Eds.). *Phages: their role in bacterial pathogenesis and biotechnology*. Washington, DC: ASM Press.

Bossi, L., Fuentes, J. A., Mora, G., and Figueroa-Bossi, N. (2003). Prophage contribution to bacterial population dynamics. *J Bacteriol*, **185**, 6467–71.

Boyd, E. F., Porwollik, S., Blackmer, F., and McClelland, M. (2003). Differences in gene content among *Salmonella enterica* serovar typhi isolates. *J Clin Microbiol*, **41**, 3823–8.

Boyd, J. S. (1950). The symbiotic bacteriophages of *Salmonella typhimurium*. *J Pathol Bacteriol*, **62**, 501–17.

Boyle, E. C., Brown, N. F., and Finlay, B. B. (2006). *Salmonella enterica* serovar Typhimurium effectors SopB, SopE, SopE2 and SipA disrupt tight junction structure and function. *Cell Microbiol*, **8**, 1946–57.

Brüssow, H., Canchaya, C., and Hardt, W. D. (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol Rev*, **68**, 560–602.

Bueno, S. M., Santiviago, C. A., Murillo, A. A., *et al.* (2004). Precise excision of the large pathogenicity island, SPI7, in *Salmonella enterica* serovar Typhi. *J Bacteriol*, **186**, 3202–13.

Campbell, A. M. (1993). Thirty years ago in *Genetics*: prophage insertion into bacterial chromosomes. *Genetics*, **133**, 433–7.

Carlson, S. A., Willson, R. M., Crane, A. J., and Ferris, K. E. (2000). Evaluation of invasion-conferring genotypes and antibiotic-induced hyperinvasive phenotypes in multiple antibiotic resistant *Salmonella typhimurium* DT104. *Microb Pathog*, **28**, 373–8.

Carlson, S. A., Meyerholz, D. K., Stabel, T. J., and Jones, B. D. (2001). Secretion of a putative cytotoxin in multiple antibiotic resistant *Salmonella enterica* serotype Typhimurium phagetype DT104. *Microb Pathog*, **31**, 201–4.

Chiu, C. H., Tang, P., Chu, C., *et al.* (2005). The genome sequence of *Salmonella enterica* serovar Choleraesuis, a highly invasive and resistant zoonotic pathogen. *Nucleic Acids Res*, **33**, 1690–8.

Collazo, C. M., and Galan, J. E. (1997). The invasion-associated type III system of *Salmonella typhimurium* directs the translocation of Sip proteins into the host cell. *Mol Microbiol*, **24**, 747–56.

Coombes, B. K., Wickham, M. E., Brown, N. F., *et al.* (2005). Genetic and molecular analysis of GogB, a phage-encoded type III-secreted substrate in *Salmonella enterica* serovar typhimurium with autonomous expression from its associated phage. *J Mol Biol*, **348**, 817–30.

Crosa, J. H., Brenner, D. J., Ewing, W. H., and Falkow, S. (1973). Molecular relationships among the Salmonellae. *J Bacteriol*, **115**, 307–15.

Crowl, R. M., Boyce, R. P., and Echols, H. (1981). Repressor cleavage as a prophage induction mechanism: hypersensitivity of a mutant lambda cI protein to *recA*-mediated proteolysis. *J Mol Biol*, **152**, 815–9.

De Groote, M. A., Ochsner, U. A., Shiloh, M. U., *et al.* (1997). Periplasmic superoxide dismutase protects *Salmonella* from products of phagocyte NADPH-oxidase and nitric oxide synthase. *Proc Natl Acad Sci USA*, **94**, 13997–4001.

Deng, W., Liou, S. R., Plunkett, G. 3rd, *et al.* (2003). Comparative genomics of Salmonella enterica serovar Typhi strains Ty2 and CT18. *J Bacteriol*, **185**, 2330–7.

Ehrbar, K., and Hardt, W. D. (2005). Bacteriophage-encoded type III effectors in *Salmonella enterica* subspecies 1 serovar Typhimurium. *Infect Genet Evol*, **5**, 1–9.

Fang, F. C., De Groote, M. A., Foster, J. W., *et al.* (1999). Virulent *Salmonella typhimurium* has two periplasmic Cu, Zn-superoxide dismutases. *Proc Natl Acad Sci USA*, **96**, 7502–7.

Fierer, J., and Guiney, D. G. (2001). Diverse virulence traits underlying different clinical outcomes of *Salmonella* infection. *J Clin Invest*, **107**, 775–80.

Figueroa-Bossi, N., and Bossi, L. (1999). Inducible prophages contribute to *Salmonella* virulence in mice. *Mol Microbiol*, **33**, 167–76.

Figueroa-Bossi, N., and Bossi, L. (2004). Resuscitation of a defective prophage in *Salmonella* cocultures. *J Bacteriol*, **186**, 4038–41.

Figueroa-Bossi, N., Coissac, E., Netter, P., and Bossi, L. (1997). Unsuspected prophage-like elements in *Salmonella typhimurium*. *Mol Microbiol*, **25**, 161–73.

Figueroa-Bossi, N., Uzzau, S., Maloriol, D., and Bossi, L. (2001). Variable assortment of prophages provides a transferable repertoire of pathogenic determinants in *Salmonella*. *Mol Microbiol*, **39**, 260–71.

Figueroa-Bossi, N., Ammendola, S., and Bossi, L. (2006). Differences in gene expression levels and in enzymatic qualities account for the uneven contribution of superoxide dismutases SodCI and SodCII to pathogenicity in *Salmonella enterica*. *Microbes Infect*, **8**, 1569–78.

Friebel, A., Ilchmann, H., Aepfelbacher, M., *et al*. (2001). SopE and SopE2 from *Salmonella typhimurium* activate different sets of RhoGTPases of the host cell. *J Biol Chem*, **276**, 34035–40.

Fukazawa, Y., and Hartman, P. E. (1964). A P22 bacteriophage mutant defective in antigen conversion. *Virology*, **23**, 279–83.

Gabbianelli, R., D'orazio, M., Pacello, F., *et al*. (2004). Distinctive functional features in prokaryotic and eukaryotic Cu,Zn superoxide dismutases. *Biol Chem*, **385**, 749–54.

Galan, J. E., and Curtiss, R. (1989). Cloning and molecular characterization of genes whose products allow *Salmonella typhimurium* to penetrate tissue culture cells. *Proc Natl Acad Sci USA*, **86**, 6383–7.

Glynn, M. K., Bopp, C., Dewitt, W., *et al*. (1998). Emergence of multidrug-resistant *Salmonella enterica* serotype typhimurium DT104 infections in the United States. *N Engl J Med*, **338**, 1333–8.

Golubeva, Y. A., and Slauch, J. M. (2006). *Salmonella enterica* serovar Typhimurium periplasmic superoxide dismutase SodCI is a member of the PhoPQ regulon and is induced in macrophages. *J Bacteriol*, **188**, 7853–61.

Groisman, E. A. (2001). The pleiotropic two-component regulatory system PhoP-PhoQ. *J Bacteriol*, **183**, 1835–42.

Groisman, E. A., and Mouslim, C. (2006). Sensing by bacterial regulatory systems in host and non-host environments. *Nat Rev Microbiol*, **4**, 705–9.

Groisman, E. A., and Ochman, H. (1996). Pathogenicity islands: bacterial evolution in quantum leaps. *Cell*, **87**, 791–4.

Gunn, J. S., Belden, W. J., and Miller, S. I. (1998). Identification of PhoP-PhoQ activated genes within a duplicated region of the *Salmonella typhimurium* chromosome. *Microb Pathog*, **25**, 77–90.

Hapfelmeier, S., Ehrbar, K., Stecher, B., *et al*. (2004). Role of the *Salmonella* pathogenicity island 1 effector proteins SipA, SopB, SopE, and SopE2 in *Salmonella enterica* subspecies 1 serovar Typhimurium colitis in streptomycin-pretreated mice. *Infect Immun*, **72**, 795–809.

Haraga, A., and Miller, S. I. (2003). A *Salmonella enterica* serovar Typhimurium translocated leucine-rich repeat effector protein inhibits NF-kappa B-dependent gene expression. *Infect Immun*, **71**, 4052–8.

Haraga, A., and Miller, S. I. (2006). A *Salmonella* type III secretion effector interacts with the mammalian serine/threonine protein kinase PKN1. *Cell Microbiol*, **8**, 837–46.

Hardt, W. D., Urlaub, H., and Galan, J. E. (1998). A substrate of the centisome 63 type III protein secretion system of *Salmonella typhimurium* is encoded by a cryptic bacteriophage. *Proc Natl Acad Sci USA*, **95**, 2574–9.

Hensel, M. (2004). Evolution of pathogenicity islands of *Salmonella enterica*. *Int J Med Microbiol*, **294**, 95–102.

Hermans, A. P., Abee, T., Zwietering, M. H., and Aarts, H. J. (2005). Identification of novel *Salmonella enterica* serovar Typhimurium DT104-specific prophage and nonprophage chromosomal sequences among serovar Typhimurium isolates by genomic subtractive hybridization. *Appl Environ Microbiol*, **71**, 4979–85.

Hermans, A. P., Beuling, A. M., Van Hoek, A. H., *et al*. (2006). Distribution of prophages and SGI-1 antibiotic-resistance genes among different *Salmonella enterica* serovar Typhimurium isolates. *Microbiology*, **152**, 2137–47.

Ho, T. D., Figueroa-Bossi, N., Wang, M., *et al*. (2002). Identification of GtgE, a novel virulence factor encoded on the Gifsy-2 bacteriophage of *Salmonella enterica* serovar Typhimurium. *J Bacteriol*, **184**, 5234–9.

Hoyer, L. L., Hamilton, A. C., Steenbergen, S. M., and Vimr, E. R. (1992). Cloning, sequencing and distribution of the *Salmonella typhimurium* LT2 sialidase gene, *nanH*, provides evidence for interspecies gene transfer. *Mol Microbiol*, **6**, 873–84.

Kingsley, R. A., and Bäumler, A. J. (2000). Host adaptation and the emergence of infectious disease: the *Salmonella* paradigm. *Mol Microbiol*, **36**, 1006–14.

Klumpp, J., and Fuchs, T. M. (2007). Identification of novel genes in genomic islands that contribute to *Salmonella typhimurium* replication in macrophages. *Microbiology*, **153**, 1207–20.

Kourilsky, P. (1973). Lysogenization by bacteriophage lambda. I. Multiple infection and the lysogenic response. *Mol Gen Genet*, **122**, 183–95.

Krishnakumar, R., Craig, M., Imlay, J. A., and Slauch, J. M. (2004). Differences in enzymatic properties allow SodCI but not SodCII to contribute to virulence in *Salmonella enterica* serovar Typhimurium strain 14028. *J Bacteriol*, **186**, 5230–8.

Krishnakumar, R., Kim, B., Mollo, E. A., Imlay, J. A., and Slauch, J. M. (2007). Structural properties of periplasmic SodCI that correlate with virulence in *Salmonella enterica* serovar Typhimurium. *J Bacteriol*, **189**, 4343–52.

Ku, Y. W., McDonough, S. P., Palaniappan, R. U., Chang, C. F., and Chang, Y. F. (2005). Novel attenuated *Salmonella enterica* serovar Choleraesuis strains as live vaccine candidates generated by signature-tagged mutagenesis. *Infect Immun*, **73**, 8194–203.

Kutsukake, K., Nakashima, H., Tominaga, A., and Abo, T. (2006). Two DNA invertases contribute to flagellar phase variation in *Salmonella enterica* serovar Typhimurium strain LT2. *J Bacteriol*, **188**, 950–7.

Lawley, T. D., Chan, K., Thompson, L. J., *et al.* (2006). Genome-wide screen for *Salmonella* genes required for long-term systemic infection of the mouse. *PLoS Pathog*, **2**, e11.

Le Minor, L. (1984). *Salmonella lignieres*. In Kreig, N. R., and Holt, J. G. (Eds.). *Bergey's manual of systematic bacteriology*. Baltimore: Williams & Wilkins.

Lemire, S. (2006). Les prophages de *Salmonella typhimurium*: Régulation lysogénique et contribution à la pathogénicité. *PhD thesis*. Paris-Sud 11 University.

Leong, J. M., Nunes-Duby, S., Lesser, C. F., *et al.* (1985). The phi 80 and P22 attachment sites. Primary structure and interaction with *Escherichia coli* integration host factor. *J Biol Chem*, **260**, 4468–77.

Matsushiro, A., Sato, K., Miyamoto, H., Yamamura, T., and Honda, T. (1999). Induction of prophages of enterohemorrhagic *Escherichia coli* O157:H7 with norfloxacin. *J Bacteriol*, **181**, 2257–60.

McClelland, M., Sanderson, K. E., Spieth, J., *et al.* (2001). Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. *Nature*, **413**, 852–6.

McClelland, M., Sanderson, K. E., Clifton, S. W., *et al.* (2004). Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of *Salmonella enterica* that cause typhoid. *Nat Genet*, **36**, 1268–74.

Miao, E. A., and Miller, S. I. (2000). A conserved amino acid sequence directing intracellular type III secretion by *Salmonella typhimurium*. *Proc Natl Acad Sci USA*, **97**, 7539–44.

Miao, E. A., Scherer, C. A., Tsolis, R. M., *et al.* (1999). *Salmonella typhimurium* leucine-rich repeat proteins are targeted to the SPI1 and SPI2 type III secretion systems. *Mol Microbiol*, **34**, 850–64.

Miao, E. A., Brittnacher, M., Haraga, A., *et al.* (2003). *Salmonella* effectors translocated across the vacuolar membrane interact with the actin cytoskeleton. *Mol Microbiol*, **48**, 401–15.

Mirold, S., Rabsch, W., Rohde, M., *et al.* (1999). Isolation of a temperate bacteriophage encoding the type III effector protein SopE from an epidemic *Salmonella typhimurium* strain. *Proc Natl Acad Sci USA*, **96**, 9845–50.

Mirold, S., Ehrbar, K., Weissmüller, A., *et al.* (2001). *Salmonella* host cell invasion emerged by acquisition of a mosaic of separate genetic elements, including *Salmonella* pathogenicity island 1 (SPI1), SPI5, and *sopE2*. *J Bacteriol*, **183**, 2348–58.

Mmolawa, P. T., Schmieger, H., and Heuzenroeder, M. W. (2003a). Bacteriophage ST64B, a genetic mosaic of genes from diverse sources isolated from *Salmonella enterica* serovar typhimurium DT 64. *J Bacteriol*, **185**, 6481–5.

Mmolawa, P. T., Schmieger, H., Tucker, C. P., and Heuzenroeder, M. W. (2003b). Genomic structure of the *Salmonella enterica* serovar Typhimurium DT 64

bacteriophage ST64T: evidence for modular genetic architecture. *J Bacteriol*, **185**, 3473–5.

Ochman, H., Soncini, F. C., Solomon, F., and Groisman, E. A. (1996). Identification of a pathogenicity island required for *Salmonella* survival in host cells. *Proc Natl Acad Sci USA*, **93**, 7800–4.

Ohl, M. E., and Miller, S. I. (2001). *Salmonella*: a model for bacterial pathogenesis. *Annu Rev Med*, **52**, 259–74.

Parkhill, J., Dougan, G., James, K. D., *et al.* (2001). Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature*, **413**, 848–52.

Pascopella, L., Raupach, B., Ghori, N., *et al.* (1995). Host restriction phenotypes of *Salmonella typhi* and *Salmonella gallinarum*. *Infect Immun*, **63**, 4329–35.

Patel, J. C., and Galan, J. E. (2006). Differential activation and function of Rho GTPases during *Salmonella*-host cell interactions. *J Cell Biol*, **175**, 453–63.

Pelludat, C., Mirold, S., and Hardt, W. D. (2003). The SopEPhi phage integrates into the *ssrA* gene of *Salmonella enterica* serovar Typhimurium A36 and is closely related to the Fels-2 prophage. *J Bacteriol*, **185**, 5182–91.

Perna, N. T., Plunkett, G., 3rd, Burland, V., *et al.* (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*, **409**, 529–33.

Plunkett, G. 3rd, Rose, D. J., Durfee, T. J., and Blattner, F. R. (1999). Sequence of Shiga toxin 2 phage 933W from *Escherichia coli* O157:H7: Shiga toxin as a phage late-gene product. *J Bacteriol*, **181**, 1767–78.

Popoff, M. Y., Bockemühl, J., and Gheesling, L. L. (2004). Supplement 2002 (no. 46) to the Kauffmann-White scheme. *Res Microbiol*, **155**, 568–70.

Porwollik, S., Boyd, E. F., Choy, C., *et al.* (2004). Characterization of *Salmonella enterica* subspecies I genovars by use of microarrays. *J Bacteriol*, **186**, 5883–98.

Prager, R., Mirold, S., Tietze, E., *et al.* (2000). Prevalence and polymorphism of genes encoding translocated effector proteins among clinical isolates of *Salmonella enterica*. *Int J Med Microbiol*, **290**, 605–17.

Rabsch, W., Andrews, H. L., Kingsley, R. A., *et al.* (2002). *Salmonella enterica* serotype Typhimurium and its host-adapted variants. *Infect Immun*, **70**, 2249–55.

Reen, F. J., Boyd, E. F., Porwollik, S., *et al.* (2005). Genomic comparisons of *Salmonella enterica* serovar Dublin, Agona, and Typhimurium strains recently isolated from milk filters and bovine samples from Ireland, using a *Salmonella* microarray. *Appl Environ Microbiol*, **71**, 1616–25.

Reeves, M. W., Evins, G. M., Heiba, A. A., Plikaytis, B. D., and Farmer, J. J. D. (1989). Clonal nature of *Salmonella typhi* and its genetic relatedness to other

salmonellae as shown by multilocus enzyme electrophoresis, and proposal of *Salmonella bongori* comb. nov. *J Clin Microbiol*, **27**, 313–20.

Roberts, J. W., and Roberts, C. W. (1975). Proteolytic cleavage of bacteriophage lambda repressor in induction. *Proc Natl Acad Sci USA*, **72**, 147–51.

Saitoh, M., Tanaka, K., Nishimori, K., *et al.* (2005). The *artAB* genes encode a putative ADP-ribosyltransferase toxin homologue associated with *Salmonella enterica* serovar Typhimurium DT104. *Microbiology*, **151**, 3089–96.

Santos, R. L., Zhang, S., Tsolis, R. M., Bäumler, A. J., and Adams, L. G. (2002). Morphologic and molecular characterization of *Salmonella typhimurium* infection in neonatal calves. *Vet Pathol*, **39**, 200–15.

Schicklmaier, P., and Schmieger, H. (1995). Frequency of generalized transducing phages in natural isolates of the *Salmonella typhimurium* complex. *Appl Environ Microbiol*, **61**, 1637–40.

Schicklmaier, P., Moser, E., Wieland, T., Rabsch, W., and Schmieger, H. (1998). A comparative study on the frequency of prophages among natural isolates of *Salmonella* and *Escherichia coli* with emphasis on generalized transducers. *Antonie Van Leeuwenhoek*, **73**, 49–54.

Schicklmaier, P., Wieland, T., and Schmieger, H. (1999). Molecular characterization and module composition of P22-related *Salmonella* phage genomes. *J Biotechnol*, **73**, 185–94.

Shea, J. E., Hensel, M., Gleeson, C., and Holden, D. W. (1996). Identification of a virulence locus encoding a second type III secretion system in *Salmonella typhimurium*. *Proc Natl Acad Sci USA*, **93**, 2593–7.

Stanley, T. L., Ellermeier, C. D., and Slauch, J. M. (2000). Tissue-specific gene expression identifies a gene in the lysogenic phage Gifsy-1 that affects *Salmonella enterica* serovar typhimurium survival in Peyer's patches. *J Bacteriol*, **182**, 4406–13.

Stender, S., Friebel, A., Linder, S., *et al.* (2000). Identification of SopE2 from *Salmonella typhimurium*, a conserved guanine nucleotide exchange factor for Cdc42 of the host cell. *Mol Microbiol*, **36**.

Susskind, M. M., and Botstein, D. (1978). Molecular genetics of bacteriophage P22. *Microbiol Rev*, **42**, 385–413.

Tanaka, K., Nishimori, K., Makino, S., *et al.* (2004). Molecular characterization of a prophage of *Salmonella enterica* serotype Typhimurium DT104. *J Clin Microbiol*, **42**, 1807–12.

Thomson, N., Baker, S., Pickard, D., *et al.* (2004). The role of prophage-like elements in the diversity of *Salmonella enterica* serovars. *J Mol Biol*, **339**, 279–300.

Threlfall, E. J., Frost, J. A., Ward, L. R., and Rowe, B. (1994). Epidemic in cattle and humans of *Salmonella typhimurium* DT 104 with chromosomally integrated multiple drug resistance. *Vet Rec*, **134**, 577.

Threlfall, E. J., Frost, J. A., Ward, L. R., and Rowe, B. (1996). Increasing spectrum of resistance in multiresistant *Salmonella typhimurium*. *Lancet*, **347**, 1053–4.

Tsolis, R. M., Townsend, S. M., Miao, E. A., *et al*. (1999). Identification of a putative *Salmonella enterica* serotype Typhimurium host range factor with homology to IpaH and YopM by signature-tagged mutagenesis. *Infect Immun*, **67**, 6385–93.

Tyler, J. S., Mills, M. J., and Friedman, D. I. (2004). The operator and early promoter region of the Shiga toxin type 2-encoding bacteriophage 933W and control of toxin expression. *J Bacteriol*, **186**, 7670–9.

Uzzau, S., Brown, D. J., Wallis, T., *et al*. (2000). Host adapted serotypes of *Salmonella enterica*. *Epidemiol Infect*, **125**, 229–55.

Uzzau, S., Figueroa-Bossi, N., Rubino, S., and Bossi, L. (2001). Epitope tagging of chromosomal genes in *Salmonella*. *Proc Natl Acad Sci USA*, **98**, 15264–9.

Uzzau, S., Bossi, L., and Figueroa-Bossi, N. (2002). Differential accumulation of *Salmonella* [Cu, Zn] superoxide dismutases SodCI and SodCII in intracellular bacteria: correlation with their relative contribution to pathogenicity. *Mol Microbiol*, **46**, 147–56.

Valdivia, R. H., and Falkow, S. (1997). Fluorescence-based isolation of bacterial genes expressed within host cells. *Science*, **277**, 2007–11.

Vander Byl, C., and Kropinski, A. M. (2000). Sequence of the genome of *Salmonella* bacteriophage P22. *J Bacteriol*, **182**, 6472–81.

Vazquez-Torres, A., and Fang, F. C. (2001). *Salmonella* evasion of the NADPH phagocyte oxidase. *Microbes Infect*, **3**, 1313–20.

Vlisidou, I., Marches, O., Dziva, F., *et al*. (2006). Identification and characterization of EspK, a type III secreted effector protein of enterohaemorrhagic *Escherichia coli* O157:H7. *FEMS Microbiol Lett*, **263**, 32–40.

Wagner, P. L., Neely, M. N., Zhang, X., *et al*. (2001). Role for a phage promoter in Shiga toxin 2 expression from a pathogenic *Escherichia coli* strain. *J Bacteriol*, **183**, 2081–5.

Wagner, P. L., Livny, J., Neely, M. N., *et al*. (2002). Bacteriophage control of Shiga toxin 1 production and release by *Escherichia coli*. *Mol Microbiol*, **44**, 957–70.

Waldor, M. K., and Friedman, D. I. (2005). Phage regulatory circuits and virulence gene expression. *Curr Opin Microbiol*, **8**, 459–65.

Wallis, T. S., and Galyov, E. E. (2000). Molecular basis of *Salmonella*-induced enteritis. *Mol Microbiol*, **36**, 997–1005.

Ward, L. R., De Sa, J. D., and Rowe, B. (1987). A phage-typing scheme for *Salmonella enteritidis*. *Epidemiol Infect*, **99**, 291–4.

Wong, K. K., McClelland, M., Stillwell, L. C., *et al.* (1998). Identification and sequence analysis of a 27-kilobase chromosomal fragment containing a *Salmonella* pathogenicity island located at 92 minutes on the chromosome map of *Salmonella enterica* serovar Typhimurium LT2. *Infect Immun*, **66**, 3365–71.

Wood, M. W., Rosqvist, R., Mullan, P. B., Edwards, M. H., and Galyov, E. E. (1996). SopE, a secreted protein of *Salmonella dublin*, is translocated into the target eukaryotic cell via a *sip*-dependent mechanism and promotes bacterial entry. *Mol Microbiol*, **22**, 327–38.

Wood, M. W., Jones, M. A., Watson, P. R., *et al.* (1998). Identification of a pathogenicity island required for *Salmonella* enteropathogenicity. *Mol Microbiol*, **29**, 883–91.

Worley, M. J., Ching, K. H., and Heffron, F. (2000). *Salmonella* SsrB activates a global regulon of horizontally acquired genes. *Mol Microbiol*, **36**, 749–61.

Worley, M. J., Nieman, G. S., Geddes, K., and Heffron, F. (2006). *Salmonella typhimurium* disseminates within its host by manipulating the motility of infected cells. *Proc Natl Acad Sci USA*, **103**, 17915–20.

Zhang, S., Santos, R. L., Tsolis, R. M., *et al.* (2002). Phage mediated horizontal transfer of the *sopE1* gene increases enteropathogenicity of *Salmonella enterica* serotype Typhimurium for calves. *FEMS Microbiol Lett*, **217**, 243–7.

Zhang, S., Kingsley, R. A., Santos, R. L., *et al.* (2003). Molecular pathogenesis of *Salmonella enterica* serotype Typhimurium-induced diarrhea. *Infect Immun*, **71**, 1–12.

Zhang, X., McDaniel, A. D., Wolf, L. E., *et al.* (2000). Quinolone antibiotics induce Shiga toxin-encoding bacteriophages, toxin production, and death in mice. *J Infect Dis*, **181**, 664–70.

Zinder, N. D., and Lederberg, J. (1952). Genetic exchange in *Salmonella*. *J Bacteriol*, **64**, 679–99.

CHAPTER 8

Pathogenic *Yersinia*: Stepwise Gain of Virulence due to Sequential Acquisition of Mobile Genetic Elements

Elisabeth Carniel

8.1. INTRODUCTION

Acquisition of genetic elements by horizontal transfer has played a major role in the evolution, virulence, and transmission of many bacteria. The genus *Yersinia* represents a very good example of a bacterial group whose pathogenicity and transmission progressively evolved with the gradual acquisition of foreign genetic elements.

Yersinia are Gram-negative bacteria that belong to the family Enterobacteriaceae. The genus is composed of 12 species that can be differentiated into pathogenic (*Y. pseudotuberculosis, Y. enterocolitica*, and *Y. pestis*) and nonpathogenic (*Y. intermedia, Y. kristensenii, Y. fredericksenii, Y. aldovae, Y. rohdei, Y. bercovieri, Y. mollaretii*, and *Y. aleksiciae*) species (Figure 8.1; see also color plate after p. 174). *Y. ruckeri* is not discussed here because the inclusion of this fish pathogen in the genus *Yersinia* is still controversial. Like several other species of *Enterobacteriacea*e, *Y. enterocolitica* and *Y. pseudotuberculosis* are true enteropathogens, while *Y. pestis* is the causative agent of plague. This chapter focuses on the evolution and lateral gene transfer in these three pathogenic species.

Y. enterocolitica and *Y. pseudotuberculosis* are widely spread in countries with temperate climates. They are transmitted by the fecal-oral route and cause intestinal symptoms such as abdominal pain (especially *Y. pseudotuberculosis*), diarrhea (especially *Y. enterocolitica*), and fever, usually of moderate intensity.

The species *Y. enterocolitica* is subdivided into six biotypes (1A, 1B, and 2 to 5). All strains are pathogenic except those of biotype 1A. Pathogenic *Y. enterocolitica* can be further subdivided into low- and high-pathogenicity strains. Low-pathogenicity *Y. enterocolitica* belong to biotypes 2 to 5. These biotypes are globally distributed and usually cause mild intestinal symptoms

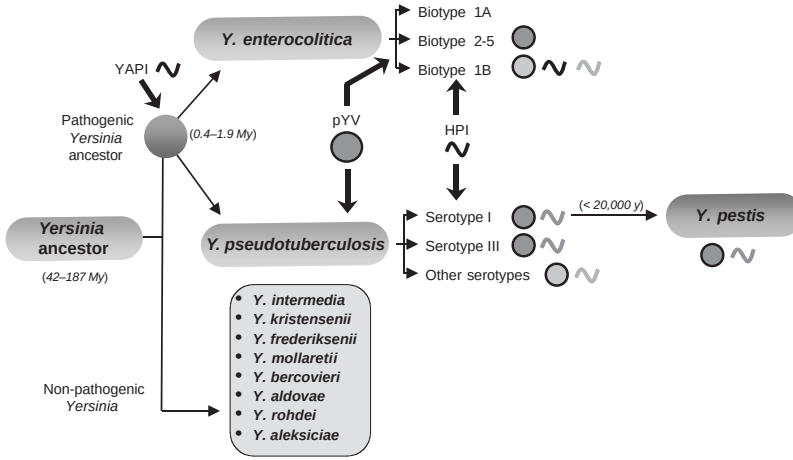


Figure 8.1. Horizontal acquisition of virulence factors by enteropathogenic *Yersinia*.

in humans. High-pathogenicity strains belong to biotype 1B, which is restricted to a few countries (e.g., the United States, Japan) and is responsible for severe systemic infections.

Y. pseudotuberculosis is subdivided into numerous serotypes and subserotypes, but serotype I largely predominates and is the most frequent cause of human pseudotuberculosis, followed by serotype III. All *Y. pseudotuberculosis* are pathogenic but serotype I strains appear to have a higher pathogenic potential than most other serotypes.

The emergence of the two enteropathogenic species was characterized by the horizontal acquisition of mobile elements conferring virulence properties. To date, three of them, one plasmid (pYV) and two pathogenicity islands (HPI and YAPI), have been studied in detail and are described herein.

8.2. LATERAL GENE TRANSFER AND EVOLUTION OF THE ENTEROPATHOGENIC *YERSINIA* SPECIES

In bacteria where horizontal genetic exchange is rare, sequence polymorphism reflects the accumulation of mutations at a uniform clock rate and correlates with the time elapsed since the existence of a last common ancestor. The use of two clock rates calibrated for *Escherichia coli* established that the *Yersinia* common ancestor arose 42 to 187 million years ago and that the two enteropathogenic species *Y. pseudotuberculosis* and *Y. enterocolitica* diverged from their common ancestor approximately 0.4 to 1.9 million years ago (Figure 8.1; Achtman *et al.*, 1999).

8.2.1. The Virulence Plasmid pYV

A conserved 70-kb plasmid, designated pYV (for **p**lasmid associated with ***Y****ersinia* **v**irulence) or alternatively pCD1 (for **p**lasmid encoding **c**alcium **d**ependence), is found uniquely in pathogenic *Yersinia* and is essential for their virulence. The roles and targets of the pYV have been extensively studied by various laboratories in recent years and are now well understood (Aepfelbacher, 2004; Aili *et al.*, 2002; Fallman and Gustavsson, 2005; Navarro *et al.*, 2005; Marenne *et al.*, 2004; Ramamurthi and Schneewind, 2002; Viboud and Bliska, 2005; Heesemann *et al.*, 2006).

8.2.1.1. pYV: A Crucial Role in *Yersinia* Pathogenicity

The major functions of this replicon are to prevent phagocytosis of the adhering bacteria and to overcome the innate and specific immune response of the mammalian host, thus allowing a better in vivo survival and multiplication of pathogenic *Yersinia*. The genes located on the pYV plasmid may be roughly divided into three major functional groups corresponding to the three main steps of interaction with the eukaryotic cells: (1) bacterial adhesion to the target cells, (2) upon contact, injection of effector proteins (Yops) into the cytosol through a needle-like structure and a pore formed in the eukaryotic cell membrane by a type III secretion machinery, and (3) alteration of various eukaryotic cell functions (disruption of the cell cytoskeleton leading to the inhibition of phagocytosis, apoptosis of macrophages, inhibition of cytokine production by T lymphocytes, reduction of surface expression of B7.2 by activated B lymphocytes, and downregulation of the inflammatory response of several cell types). During these different steps, the bacteria are in close contact with the target cell but remain extracellular.

8.2.1.2. pYV: A Plasmid of Unknown Origin

The bacterial donor from which pathogenic *Yersinia* acquired the pYV is unknown. No plasmid closely related to the entire pYV has been identified in other bacterial species. This replicon is not transferable and does not carry any sequence involved in plasmid conjugation or mobilization. Its origin of replication is homologous to that of IncFIIA plasmids. The G+C content of pYV of 44.8% is slightly lower than that of the core genome (47–48%) and is heterogeneous throughout the plasmid. This, in addition to the presence of several partial or complete insertion sequences and a different overall organization among the pathogenic *Yersinia*, suggests a complex history of DNA acquisition, deletions, insertions, translocations, and internal rearrangements (Perry *et al.*, 1998; Hu *et al.*, 1998).

The pYV-encoded type III secretion apparatus (T3SS) is similar to several other T3SSs identified in bacterial species as diverse as *Pseudomonas aeruginosa, Photorhabdus luminescens*, or *Bordetella pertussis*. Several components of T3SS have homologs in flagellae and both structures export proteins, suggesting that they are derived from a common ancestor. The fact that the evolutionary tree of T3SS is strikingly different from the phylogeny of their bacterial hosts and that these systems may be found on mobile elements such as plasmids and pathogenicity islands strongly argues for their acquisition by lateral gene transfer (Foultier *et al.*, 2002). However, except for the T3SS region, the remaining part of the pYV and more specifically the *yop* genes encoding the T3SS effectors are unique to the genus *Yersinia*.

8.2.1.3. pYV: Acquisition by Enteropathogenic *Yersinia* Species

The presence of a conserved pYV in both *Y. enterocolitica* and *Y. pseudotuberculosis* may suggest that this plasmid has been acquired vertically from the common ancestor and has been subsequently lost by the non-pathogenic group of *Y. enterocolitica* biotype 1A. However, despite a preserved gene pool, the genetic organization of the plasmid is different in the two species. Furthermore, the identity of the *Y. enterocolitica* and *Y. pseudotuberculosis* pYV-borne genes (91–99% nt identity for most coding sequences) is much higher than that of housekeeping genes (59–84% identity) acquired vertically from the common ancestor (Carniel, 2003). Together, these data rather support the hypothesis of an independent acquisition of the pYV by the *Y. enterocolitica* and *Y. pseudotuberculosis* branches, after they diverged from their common ancestor (Figure 8.1).

8.2.2. Pathogenicity Islands

Pathogenicity islands (PAI) represent a specific group of mobile elements that are found inserted in the chromosome of numerous bacterial species and confer specific pathogenic properties on the host strains (Hacker *et al.*, 1997). These elements have some characteristics that make them specific entities and which allow them to be differentiated from the surrounding chromosomal genes: They are characteristically inserted into a tRNA gene, they are flanked by short direct repeats, their G+C content is different from that of the core genome, they encode a phage-like integrase, and they have the potential to spontaneously excise from the bacterial chromosome.

Two PAI have been identified and studied in enteropathogenic *Yersinia*: the High-Pathogenicity Island (HPI) and the *Yersinia* Adhesion Pathogenicity Island (YAPI). Both islands have characteristic features of PAI.

8.2.2.1. The High-Pathogenicity Island

The HPI was the first PAI identified in *Yersinia*. The structure, function, and mobility of this island have been investigated in detail (Lesic and Carniel, 2004; Schubert *et al.*, 2004b). The HPI was named after the observation that this element is strictly associated with subgroups of *Yersinia* strains that have the potential to cause severe and systemic infections in humans (*Y. enterocolitica* 1B, *Y. pseudotuberculosis* serotype I ([and III]). The HPI carries genes encoding the biosynthesis, transport, and regulation of an iron-chelating compound designated yersiniabactin, a 482-Da molecule that belongs to a small sub-group of phenolate siderophores. In mammals, iron is bound to eukaryotic proteins which maintain a level of free iron that is far too low to sustain bacterial growth. The extremely high affinity of yersiniabactin for ferric iron allows it to capture the metal bound to host proteins and to allow uptake by the bacteria, which use it for their metabolism and for their dissemination in vivo.

8.2.2.2 HPI: Genetic Organization

The HPI is a genetic element of 36 kb (*Y. pseudotuberculosis*) to 43 kb (*Y. enterocolitica*) inserted into an *asn* tRNA site on the bacterial chromosome. The island is composed of two distinct regions: (1) a highly conserved 29-kb part involved in yersiniabactin biosynthesis and transport (referred as the *ybt* locus), and (2) a variable A+T-rich part that contains up to four insertion sequences (IS) of the IS*3* family as well as remnants of phage genes. The latter region is completely different in size and composition in *Y. enterocolitica* 1B and *Y. pseudotuberculosis* I and is missing (along with the two adjacent *ybt* genes) in *Y. pseudotuberculosis* III.

Both *Y. enterocolitica* and *Y. pseudotuberculosis* HPIs are flanked by 17-bp repeats designated *attB* sites and carry, adjacent to the *asn* tRNA insertion site, an integrase gene (*int*) homologous to that of bacteriophage P4. *int* is intact and functional in a subset of *Y. pseudotuberculosis* strains but is mutated in the other strains and in all *Y. enterocolitica* 1B.

8.2.2.3. HPI: Intra-genomic Mobility

The HPI is stabilized in the genome of *Y. enterocolitica* because of a mutation in the HPI-borne integrase gene. Remarkably, the *Y. pseudotuberculosis* HPI is mobile within the genome of its host strain: it can excise from the bacterial chromosome and reinsert into another *asn* tRNA locus on the same chromosome.

Spontaneous excision of the *Y. pseudotuberculosis* HPI occurs at a frequency of 10^{-5} to 10^{-7} by site specific recombination between the two 17-bp *attB* flanking sites. The excision process requires the combined action of the HPI-borne integrase and Hef (for HPI excision factor), a recombination directionality factor that is assumed to drive the function of Int toward an excisionase activity. A circular HPI molecule is generated upon excision. Integration of the HPI into the bacterial chromosome is mediated by Int and results in a duplication of *attB* at each extremity of the island.

8.2.2.4. HPI: A Widely Distributed PAI

To date, the HPI is the only PAI found in a wide variety of bacterial genera. After having been originally identified in the three pathogenic *Yersinia* species, this island was subsequently detected in *E. coli,* and then in various members of the family Enterobacteriaceae, such as *Enterobacter* (*E. cloacae* and *E. aerogenes*), *Klebsiella* (*K. pneumoniae, K. rhinoscleromatis, K. ozaenae, K. planticola*, and *K. oxytoca*), *Citrobacter* (*C. diversus* and *C. koseri*), *Salmonella enterica, Photorhabdus luminescens,* and *Serratia liquefaciens*. The overall organization of the island is well conserved in most of the different HPI-positive enterobacteria and most of them are able to produce yersiniabactin.

In *E. coli*, the HPI has been identified in a wide variety of pathotypes and is associated with the most severe forms of infections. As in yersiniae, this island appears to be a key factor that determines the potential of *E. coli* strains to disseminate in the host and cause systemic infections.

The *E. coli* HPI is genetically more closely related to the *Y. pseudotuberculosis* than to the *Y. enterocolitica* island. The *ybt* locus is highly conserved and the identity between the products of most *E. coli* and *Y. pseudotuberculosis* paralogs is extremely high (98–100%), suggesting a recent acquisition of the island by these bacteria.

8.2.2.5. HPI: Horizontal Transfer

The fact that the HPI has kept its mobility and is highly conserved in various bacterial genera suggests that the island has been acquired recently and may have retained its ability to be horizontally transmitted to new bacterial hosts.

PAI are considered as horizontally transferable elements but their mechanism of transmission is not yet elucidated. The fact that these islands are inserted into tRNA genes and use a phage-like machinery for integration and excision suggests that they are of phage origin. Two PAI have been shown to have the capacity to be transduced by helper phages: SaPI1 of *Staphylococcus*

aureus, which uses the helper phages 80α for its transfer, and VPI of *Vibrio cholerae*, which is transduced by CP-T1.

Recently, a model of PAI transfer has been proposed and experimentally demonstrated for the HPI (Antonenka *et al.*, 2005). In this model, the island first integrates into a specific *attB* site carried by a conjugative plasmid, which serves as a shuttle to transfer the HPI to a new bacterial host. In the recipient, the HPI excises itself from the replicon and integrates into an *attB* site on the bacterial chromosome.

Another model of HPI transfer derives from the identification of a particular *E. coli* strain (ECOR31), found to harbor an island with an additional 35-kb fragment carrying genes involved in conjugative DNA transfer (Schubert *et al.*, 2004a). This element, whose organization resembles that of an integrative and conjugative element (ICE), has been proposed to be the progenitor of the transferable HPI. Nevertheless, the fact that this additional region is absent from all other HPI-positive strains and that the putative ICE is inserted into a locus different from that of the HPI on the chromosome of other *E. coli* strains suggests that this element may represent a particular type of HPI rather than the progenitor of all HPI.

All the experimental demonstrations of PAI transfer were obtained after the artificial introduction of helper phages or conjugative functions into the tested strains. The horizontal transfer of the *Y. pseudotuberculosis* HPI to new strains without introducing exogenous transfer functions was recently observed (Lesic and Carniel, 2005). Transmission of the HPI from a *Y. pseudotuberculosis* donor to a *Y. pseudotuberculosis* or a *Y. pestis* recipient was obtained under specific conditions (low temperature, iron-deprived liquid medium) and occurred at low frequencies ($\approx 10^{-8}$). However, this lateral transfer did not require the HPI-borne excision/integration machinery and was not restricted to the HPI, but occurred by homologous recombination between the HPI flanking regions. In distantly related bacteria with non-conserved genetic organizations and low nucleotide identities, a transfer based on homologous recombination may be of poor efficacy, but this mechanism may represent a means of propagating the HPI in closely related species.

8.2.2.6. HPI: Origin and Acquisition by Highly Pathogenic *Yersinia*

The high degree of conservation between HPI-borne genes of *Yersinia* and *E. coli* argues for a recent and horizontal acquisition of the island in these two bacterial genera, long after their divergence from a common ancestor. The bacterial species that may have been the donor of the HPI is still not known.

In *Yersinia*, the HPI is systematically absent from all non-pathogenic species, suggesting that it was specifically acquired by the pathogenic branch. However, at least three pieces of evidence argue for an independent acquisition of the island by *Y. enterocolitica* and *Y. pseudotuberculosis* branches, after their divergence from the *Yersinia* common ancestor (Figure 8.1): (1) Within the enteropathogenic species, the island is restricted to subgroups of strains (biotype 1B of *Y. enterocolitica* and serotypes I and III of *Y. pseudotuberculosis*), (2) the HPI has a different genetic organization in the two species, and (3) the identity between the HPI-borne genes of *Y. enterocolitica* and *Y. pseudotuberculosis* is much higher than that of chromosomal housekeeping genes, suggesting a recent acquisition.

8.2.2.7. YAPI: An Adhesion Element

A second PAI has been recently identified in enteropathogenic *Yersinia* and has been termed YAPI for *Yersinia* adhesion pathogenicity island (Collyn *et al.*, 2004).

YAPI is a 98-kb element inserted at a *phe* tRNA locus and was identified in *Y. pseudotuberculosis*. This PAI encodes a functional type IV pilus that promotes bacterial adhesion to the intestinal mucosa and is homologous to a type IV pilus encoded by a PAI of *Salmonella enterica* serovar Typhi (SPI-7). The *pil* locus appears to be the only DNA region on YAPI that plays a role in *Y. pseudotuberculosis* pathogenicity. The 36-kb region comprising the *pil* locus is well conserved among the enteropathogenic YAPI-positive *Yersinia*. Besides the type IV pilus locus, YAPI carries a type I restriction/modification system, several genes found on mobile elements (insertion sequences, plasmids, bacteriophages), and a region similar to a *Ralstonia solanacerum* megaplasmid encoding metabolic genes. The last region is not present on all YAPI of *Y. pseudotuberculosis*.

8.2.2.8. YAPI: A Mobile Element

The island carries the sequences necessary for its excision: two 54-bp *attB* flanking sites and an integrase gene homologous to prophage integrases of the Cre family. YAPI has the capacity to excise precisely from the bacterial chromosome of its host strain with a frequency of 10^{-7}. The observation that YAPI may be inserted into either of the two *phe* tRNA loci present on the *Y. pseudotuberculosis* chromosome may indicate either an independent acquisition of the island by different strains or a potential of intragenomic mobility (as observed for the HPI).

8.2.2.9. YAPI: Origin and Acquisition by Highly Pathogenic *Yersinia*

YAPI is not present in all *Y. pseudotuberculosis* strains but is found mostly in Asian isolates, independently of their O-serotype (Collyn *et al.*, 2005). Its distribution among non-pathogenic *Yersinia* is not known. Although the presence of YAPI in the species *Y. enterocolitica* has not yet been studied, this PAI has been identified on the sequenced chromosome of *Y. enterocolitica* 1B strain 8081. The YAPI region encompassing the *pil* locus is conserved in *Y. pseudotuberculosis* and *Y. enterocolitica*, but the other part of the island is different in the two species. The facts that YAPI has a heterogeneous G+C content and that one portion of the island is either missing or differing among various strains reflects a chimerical genetic structure that may have resulted from several insertion, excision, and recombination events.

The existence of various bacterial mobile elements carrying a conserved *pil* locus combined with the analysis of nucleotide adaptation and gene phylogeny recently led to the proposal of the following evolutionary scenario: The ancestral YAPI arose from the integration of a plasmid of the R64 family carrying the *pil* locus in the *Salmonella* chromosome and this new island (ancestor of SPI-7) secondarily spread to other enterobacteria such as *Yersinia* and *Photorhabdus* (Collyn *et al.*, 2006).

The observations that the distances separating the *Y. enterocolitica* and *Y. pseudotuberculosis* YAPIs from the other *pil*-harboring genetic elements are identical and that the YAPI-borne and chromosomal genes share the same degree of identity in the two enteropathogenic *Yersinia* species argue for the acquisition of the island by the *Y. enterocolitica* and *Y. pseudotuberculosis* common ancestor (Figure 8.1; Collyn *et al.*, 2006). If so, this unstable island has been secondarily lost by some subsets of *Y. pseudotuberculosis* strains.

8.2.3. A Model for the Evolution of Enteropathogenic *Yersinia*

The sequential acquisition of two PAIs and a plasmid has been essential for the emergence of the enteropathogenic *Yersinia* (Figure 8.1). YAPI might have been acquired first by the common ancestor and then transmitted vertically to the two enteropathogenic species. It is possible that its maintenance in some subsets of *Y. pseudotuberculosis* from Asia is related to the specific clinical manifestations observed in Russia and Japan (although this relation deserves further demonstration). Current evidences suggest that the pYV and the HPI were acquired secondarily and independently by the *Y. enterocolitica* and *Y. pseudotuberculosis* branches. Based on their respective distribution and conservation, it may be hypothesized that the pYV was acquired before the HPI (Figure 8.1). While horizontal acquisition of the pYV has obviously been

crucial for the transformation of environmental bacteria into pathogens, the secondary uptake of the HPI by some pathogenic *Y. enterocolitica* and *Y. pseudotuberculosis* has led to their transformation into bacteria able to disseminate in their host and cause systemic infections. Therefore, in enteropathogenic *Yersinia* the gradual gain in pathogenicity resulted from the sequential acquisition of these mobile elements.

Once more genomes of various *Yersinia* species are available, comparative genomics followed by functional studies will likely allow delineation of additional pathogenicity-related chromosomal regions horizontally acquired by the enteropathogenic group of *Yersinia*.

8.3. LATERAL GENE TRANSFER AND EVOLUTION OF *Y. PESTIS*

Y. pestis, the causative agent of plague, has been one of the most devastating infectious agents in world history. During the Christian era, three well-documented plague pandemics occurred: Justinian's plague (sixth and seventh centuries); the "Black Death" (Middle Ages to 18th century); and the third pandemic, which started from Hong Kong in 1894. Endemic plague foci persist nowadays in Africa, Asia, and North and South America, and the plague is currently categorized as a reemerging disease.

Y. pestis has clinical and epidemiological features drastically different from those of *Y. enterocolitica* and *Y. pseudotuberculosis*. This bacillus is found predominantly in tropical and subtropical areas, it is transmitted by flea bites, it has a restricted rodent-flea-rodent cycle, and it causes a highly severe and often fatal disease. While infections due to enteropathogenic *Yersinia* are most often mild and self-limiting, bubonic plague is lethal in 40% to 70% of the cases, in usually less than a week, and pneumonic plague is systematically fatal, most often in 3 days, if an effective antibiotherapy is not administered early.

8.3.1. Evolution and Microevolution of *Y. pestis*

8.3.1.1. *Y. pestis*, a Clone of *Y. pseudotuberculosis*

Strikingly, despite their dramatically dissimilar life cycles and pathogenicity potentials, DNA-DNA hybridizations experiments performed in the 1980s showed that *Y. pestis* and *Y. pseudotuberculosis* are genetically almost identical (>90% chromosomal relatedness; Bercovier *et al.*, 1980). These bacteria were nonetheless maintained as different species to avoid any risk of confusion for clinical laboratories.

This high genetic relatedness was further confirmed recently by a multilocus sequence typing (MLST) analysis performed on a large number of *Y. pestis*, *Y. pseudotuberculosis*, and *Y. enterocolitica* isolates. This analysis demonstrated a complete absence of variation at 21,881 synonymous sites in *Y. pestis*, indicating that this species is highly clonal (Achtman *et al.*, 1999). Remarkably, the distances between the alleles in *Y. pestis* and *Y. pseudotuberculosis* were within the range seen in *Y. pseudotuberculosis* alone, indicating that *Y. pestis* is a highly conserved clone of *Y. pseudotuberculosis*. Unexpectedly, the use of *E. coli* molecular clock rates indicated that *Y. pestis* is a very recent clone, which emerged from *Y. pseudotuberculosis* within the past 1,500 to 20,000 years.

None of the *Y. pseudotuberculosis* strains analyzed had all alleles identical to those of *Y. pestis* and therefore none of them could be considered as the direct progenitor of the plague bacillus. However, two pieces of evidence suggest that the ancestor was a strain of serotype I. The first piece of evidence came from the analysis of the antigen O locus, which is polymorphic and defines the serotypes in *Y. pseudotuberculosis*. Strains of serotype I had the genetic organization of their O-antigen clusters most similar to that of *Y. pestis* (Skurnik *et al.*, 2000). The second piece of evidence came from the analysis of the HPI, whose organization and nucleotide sequence are almost 100% identical over its entire length in *Y. pestis* and *Y. pseudotuberculosis* of serotype I.

The recent emergence of *Y. pestis* along with the high conservation of the pYV and the HPI in both *Y. pestis* and *Y. pseudotuberculosis* indicate that these two mobile elements were most likely acquired vertically by the plague bacillus from its *Y. pseudotuberculosis* progenitor.

8.3.1.2. Microevolution of *Y. pestis*

As a young pathogen, *Y. pestis* has evolved too recently to allow the accumulation of extensive sequence diversity. However, this species is not totally uniform and is subdivided into four biovars based on biochemical traits: Pestoides (or Microtus), Antiqua, Medievalis, and Orientalis. The use of a combination of three different multilocus molecular methods targeting genome-wide synonymous single-nucleotide polymorphisms (SNPs), variation in size of tandem repeats, and insertion of IS*100* elements allowed the construction of a *Y. pestis* microevolutionary tree rooted on *Y. pseudotuberculosis* (Achtman *et al.*, 2004). This tree (Figure 8.2; see also color plate after p. 174) indicates that *Y. pestis* initially evolved from *Y. pseudotuberculosis* along one branch, called branch 0, from which Pestoides split off. Branch 0 was followed by a binary split into branch 1 and branch 2, approximately 6,500 years ago.

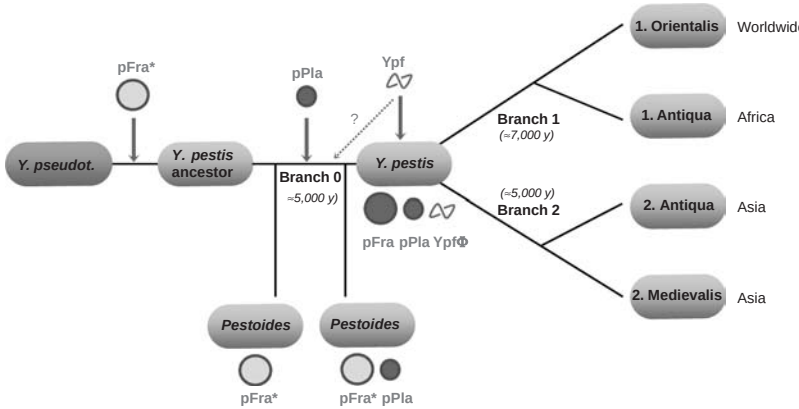


Figure 8.2. Microevolution of *Y. pestis*. pFra*: ancestor of pFra, with a larger size and conjugation genes.

Branch 1 includes Orientalis isolates of worldwide origin and Antiqua strains from Africa, while branch 2 comprises Medievalis and Antiqua strains from Asia (Figure 8.2). The recent availability of two additional *Y. pestis* genomes of biovar Antiqua (Chain *et al.*, 2006) further strengthens the validity of the *Y. pestis* microevolutionary tree.

High diversity is often a good indicator of the geographical source of microbes. The isolation of representatives of branches 0, 1, and 2 from Asia suggests that *Y. pestis* arose in this part of the world rather than in Africa, from which branch 2 has not been isolated.

8.3.1.3. *Y. pestis* Has Very Few Genetic Differences with *Y. pseudotuberculosis*

The comparisons of the genome sequences of one *Y. pseudotuberculosis* (Chain *et al.*, 2004) and three *Y. pestis* isolates (Parkhill *et al.*, 2001; Deng *et al.*, 2002; Song *et al.*, 2004) identified genetic differences that distinguish the two species. As expected, most of the genome backbone in *Y. pestis* and *Y. pseudotuberculosis* is highly conserved, with 75% of predicted proteins having greater than 97% identity in the two species (Chain *et al.*, 2004).

One notable difference between the *Y. pestis* and *Y. pseudotuberculosis* chromosomes is a burst of insertion sequence (IS) elements in the former. Although the mechanism underlying this IS expansion is unknown, it most likely participates to the high genome plasticity of the plague bacillus and could potentially facilitate its adaptation to a variety of new ecological niches.

Remarkably, the *Y. pestis* genome is characterized by a massive loss of genetic material (317 putative genes) and gene functions (208 pseudogenes), leading to as many as 13% of *Y. pseudotuberculosis* coding sequences that are no longer functional in *Y. pestis*. It has been proposed that this process may represent an intermediate stage in genome compaction, a feature commonly observed in the evolution of pathogens closely associated with their hosts.

In contrast, the emergence of *Y. pestis* has been characterized by the acquisition of very few new genetic elements. The analysis of chromosomal gene differences across a panel of *Y. pseudotuberculosis* and *Y. pestis* isolates from around the world revealed that only seven regions and three solitary coding sequences have been acquired by *Y. pestis* since its divergence from *Y. pseudotuberculosis* (Chain *et al.*, 2004). Most of these *Y. pestis*-specific chromosomal regions have no characteristics of mobile elements, with the exception of two prophage regions that correspond to genomes of a 46.3-kb putative lambdoid defective bacteriophage and an 8.7-kb filamentous prophage. *Y. pestis* emergence has also been characterized by the acquisition of two specific plasmids designated pFra and pPla (Lindler, 2004).

Of these newly and horizontally acquired regions, only three have been studied in detail: the two *Y. pestis*-specific plasmids and the filamentous phage. These mobile genetic elements, which might have played a crucial role in evolution, new life cycle, and/or increased pathogenicity of *Y. pestis*, are described next.

8.3.2. Acquisition of pFra by *Y. pestis*

8.3.2.1. pFra: A Plasmid Important for Flea-borne Transmission of *Y. pestis*

pFra is classically a 96- to 101-kb plasmid also called pMT1. The genetic organization and size of this replicon are not well conserved within the *Y. pestis* species (Lindler, 2004). Strains deprived of the entire plasmid are still fully virulent for mice and for African green monkeys, suggesting that pFra is not necessary for the virulence of the plague bacillus. This plasmid carries about 115 putative coding sequences, of which 38% have no significant homology to any protein present in the databases. Some of the other open reading frames are homologous to genes involved in protein and DNA metabolism or regulation. Numerous mobility genes (integrases, transposases, resolvases, IS, and phage sequences) are also present. To date, only two pFra-borne loci have been studied in details: *ymt* encoding a phospholipase D (previously known as the murine toxin) and *caf*, coding for the fraction 1 antigen (F1 Ag).

The F1 Ag forms a capsule-like structure surrounding bacteria grown at 37°C. The *caf* locus is composed of four genes that encode a chaperone, a protein involved in capsule anchoring, and the capsular subunits, as well as an activator of the AraC family. The capsular subunits assemble to form large polymers at the bacterial surface (MacIntyre *et al.*, 2004). Because of its high immunogenicity, this antigen is used for plague diagnosis (enzyme-linked immunosorbent assay [ELISA], dipsticks) and is included in many vaccine preparations. However, although the presence of the F1 Ag has been associated with resistance to phagocytosis by monocytes, the role of this antigen in *Y. pestis* virulence is still a matter of debate.

The second best characterized locus, *ymt*, codes for a cytosolic phospholipase D synthesized at 26°C (Hinnebusch, 2004). This protein is not a virulence factor for mammals, but it promotes the survival of *Y. pestis* in the flea gut and the colonization of its proventriculus. The acquisition of a plasmid encoding this phospholipase may thus have been a key step in the transformation of *Y. pestis* from an enteric bacterium to a pathogen transmitted by fleas.

8.3.2.2. pFra: Origin, Acquisition, and Evolution

pFra origin of replication and partitioning functions resemble those of bacteriophages P1 and P7. This plasmid has a G+C content (50.2%) close to that of the *Y. pestis* core genome but harbors some regions with a much lower G+C content (38–39%), suggesting a mosaic structure composed of different genetic elements acquired sequentially. Plasmid sequence comparisons also indicate that numerous rearrangements took place in pFra from different *Y. pestis* isolates.

Recently, a clinical isolate of *Salmonella enterica* serovar Typhi carrying a plasmid (pHCM2) that had more than half of its sequence (62 open reading frames [ORFs]) in common with pFra was isolated in Vietnam (Prentice *et al.*, 2001). Most of the common genes were highly identical (>96% nt identity). Interestingly, pHCM2 had a higher number of common sequences with a pFra plasmid from a Medievalis strain (KIM5) than with one from an Orientalis strain (CO92), indicating that the plasmid underwent further rearrangements in the latter biovar. Furthermore, the GC content of the regions common to both plasmids is much closer to the *S. enterica* than to the *Y. pestis* chromosome, suggesting that if a transfer of the plasmid between the two species occurred, it should have been from *Salmonella* to *Y. pestis*.

pFra plasmids from Pestoides strains (branch 0, Figure 8.2) usually have a larger size than those isolated from the classical *Y. pestis* biovars. Sequence analysis of a 137-kb plasmid (pG8786) from such a strain revealed that, in

addition to the genes found in other pFra, it carries two additional regions. The first region (4.6 kb) is found and highly conserved on the pHCM2 plasmid of *S. enterica*. The second region contains 25 ORFs that are similar to transfer genes found in F-like plasmids (Golubov *et al.*, 2004). However, some of the genes necessary for conjugative transfer are missing in this region, and pG8786 is not transmissible in vitro. This observation nonetheless supports the hypothesis that the ancestral pFra carried the machinery necessary for its horizontal transmission and thus argues for its acquisition by conjugative transfer, perhaps from a *S. enterica* donor.

8.3.3. Acquisition of pPla by *Y. pestis*

8.3.3.1. pPla: A Plasmid Allowing *Y. pestis* to Disseminate from the Site of the Flea Bite

The second *Y. pestis*-specific plasmid is pPla (alternatively termed pPCP1 or pPst). This 9.6-kb replicon carries nine putative ORFs, of which five have been shown to encode an IS*100* transposase (two overlapping ORFs), a plasminogen activator (Pla), a bacteriocin called pesticin (Pst), and its immunity protein (Pim).

Pesticin, a 39.9-kDa bacteriocin, is an *N*-acetyl glucosaminidase that exerts its action on the peptidoglycan layer in the bacterial periplasm. This bacteriocin is active only on microorganisms that carry the HPI because the pesticin outer membrane receptor is encoded by the island. This bacteriocin may help the bacterium survive and compete with other organisms present in the same ecological niches, but it does not play a role in *Y. pestis* virulence. Adjacent to the *pst* gene, the *pim* gene encodes the pesticin immunity protein that protects the bacteria from the action of its own bacteriocin.

The plasminogen activator Pla is an outer membrane protein that belongs to the omptin family of aspartic proteases and has several activities in vitro (Korhonen *et al.*, 2004): (1) It increases local production of plasmin by plasminogen activation and α2-antiplasmin inactivation, (2) it degrades the complement component C3, (3) it mediates invasion of human endothelial and epithelial cells, and (4) it promotes high-affinity adhesion to basement membranes. This protein has no role in flea blockage or transmission, but it seems to be essential to overcome the lymphatic tissue barrier in bubonic plague and to shorten the retention time in the splenic filter during septicemic plague (Sebbane *et al.*, 2006).

Although some strains lacking pPla are severely attenuated (1 million-fold) after subcutaneous inoculation (Sodeinde *et al.*, 1992), some isolates conserve their full virulence, suggesting that alternative factors may serve the

function of Pla in these strains. However, the observation that two distinct pathologies can ensue from a flea bite – bubonic plague (the most frequent) and more rarely primary septicemic plague – and that Pla is required for the development of bubonic plague only (Sebbane *et al.*, 2006) may also indicate that some *Y. pestis* isolates naturally deprived of pPla can still cause septicemic plague.

8.3.3.2. pPla: Origin and Acquisition

pPla has a ColE1-like origin of replication and a G+C content of 45.3%. This plasmid has not been isolated in bacteria other than *Y. pestis*, but Pla homologs are found in various enterobacteria. The closest protein (78% identity) is PlaA, which is encoded by a 36-kb plasmid in *Erwinia pyrifolia*. Interestingly, most of the Pla homologs are encoded by mobile genetic elements such as plasmids and cryptic phages, suggesting that they spread among various *Enterobacteriaceae* by lateral gene transfer. However, no conjugation or mobilization functions have been identified on pPla.

Acquisition of pPla by a *Y. pestis* ancestor has certainly been important to facilitate the dissemination of the bacteria from the flea bite site of inoculation to the draining lymph node and then to the organs and the blood of the infected host. However, introduction of pPla in *Y. pseudotuberculosis* does not increase its pathogenicity (Pouillot *et al.*, 2005; Kutyrev *et al.*, 1999), suggesting that acquisition of Pla alone was not sufficient to cause a dramatic rise in *Y. pestis* pathogenicity.

8.3.4. Acquisition of a Filamentous Phage by *Y. pestis*

8.3.4.1. YpfΦ is a Functional Filamentous Phage

Recent comparative genomics studies have identified a few chromosomal regions acquired by *Y. pestis* during or after its divergence from *Y. pseudotuberculosis*. One of these regions is homologous to a filamentous prophage and has been designated YpfΦ (for *Y. pestis* filamentous phage; Derbise *et al.*, 2007). This 8.7-kb phage genome is inserted into the chromosomal *dif* site, a recombinational locus that functions in the resolution of chromosome dimers. In the *Y. pestis* Orientalis chromosome, the YpfΦ genome is arranged as a tandem structure composed of at least two to four directly repeated copies of the phage genome. The prophages region contains 13 putative ORFs, mainly involved in phage replication, encapsidation, and secretion. However, two ORFs of unknown functions are located at each extremity of the prophage.

YpfΦ is still functional. Filamentous phage particles that contain a circular positive single-strand DNA molecule are produced and released in the bacterial supernatant, without lysing the bacterial host. The secreted virions are infectious for other *Y. pestis* strains.

YpfΦ is not necessary for the plague bacillus to infect and block the flea *Xenopsylla cheopis*, arguing against a role of this phage in *Y. pestis* flea-borne transmission. In contrast, the presence of the prophage confers an advantage on the host strain during growth in vivo, suggesting that it participates in the infectious process by a mechanism not yet characterized.

8.3.4.2. YpfΦ Origin and Adaptation in *Y. pestis*

YpfΦ is highly similar (99% nt identity) to a prophage designated CUS-1, identified in the highly virulent *Escherichia coli* clone O18:K1:H7. The identity extends over the 7.1-kb right-hand portion of the sequence, but the two prophages differ at their left-hand extremities, where the two last ORFs are specific for each species. This extremely high sequence identity in two distantly related species as well as the existence of a few species-specific regions suggest- that both organisms were infected recently by a similar (but different) virion. The fact that the sequences specific for each species (and not involved in phage physiology) are located at the extremities of the prophages, in the vicinity of the integration site, also suggests that they have been acquired from an unknown bacterial donor strain by a mechanism of imprecise phage excision.

Within the species *Y. pestis*, YpfΦ is stably integrated into the bacterial chromosome of biovar Orientalis strains but is present as an extrachromosomal and highly unstable replicon in Antiqua and Medievalis isolates. Its maintenance in all *Y. pestis* sub-branches for more than 7500 years, despite its high in vitro instability, suggests that this mobile element has been subjected to a strong positive selective pressure.

The advantage brought by YpfΦ may be an enhanced capacity of the host strain to cause an infection. Furthermore, recent investigations of ancient DNA from dental pulp extracts collected from various mass graves suggest that the three plague pandemics were all caused by biovar Orientalis (Drancourt *et al.*, 2004). If so, it can also be speculated that the stabilization of the phage genome in the Orientalis sub-branch conferred a higher pandemic potential on this biovar.

Most likely, YpfΦ was acquired as an unstable episome by *Y. pestis* before the split of the species into branches 1 and 2 (Figure 8.2), and the phage secondarily stabilized in the 1.ORI sub-branch, upon integration of its genome into the bacterial chromosome.

8.4. AN EVOLUTIONARY SCENARIO FOR THE EMERGENCE OF *Y. PESTIS*

8.4.1. Emergence of a Pathogen Transmitted by Fleas

Because *Y. pestis* is a very recent descendant of *Y. pseudotuberculosis*, one may wonder how this organism lost the classical fecal-oral route of transmission and acquired the ability to be transmitted by fleas. A putative evolutionary scenario for the stepwise acquisition of this mode of transmission is presented next.

In this (hypothetical) scenario, the first step was the co-infection of a rodent in Asia with a *Y. pseudotuberculosis* strain of serotype I and another enterobacterial species (maybe *Salmonella enterica* serovar Typhi) which harbored the pFra ancestral molecule. The close contact between the two bacteria in the intestinal tract of the infected rodent resulted in the conjugative transfer of this replicon to *Y. pseudotuberculosis*. Since *Y. pseudotuberculosis* frequently causes septicemia in rodents, the recombinant strain reached the bloodstream, where it was taken up by rodent fleas during their blood meal on the infected animal. The presence of the ancestral pFra (carrying the *ymt* locus) allowed this new variant of *Y. pseudotuberculosis* to survive in and to colonize the flea proventriculus.

Another important step was the vertical acquisition by *Y. pestis* from its *Y. pseudotuberculosis* ancestor of genes responsible for flea blockage. Indeed, efficient transmission of *Y. pestis* to a new host by fleas requires that the bacteria form a solid mass that blocks the proventriculus of the insect. During repeated attempts to feed on this new host, the hungry blocked flea is unable to pump blood and is forced to regurgitate the bacteria into the bite wound. This sophisticated and highly efficient mechanism of flea-borne transmission developed by *Y. pestis* is encoded by the chromosomal hemin storage locus (*hms*) and is due to the formation of a biofilm in the insect gut (Hinnebusch *et al.*, 1996). Another chromosomal locus, *gmhA*, has also been recently shown to be required to block fleas (Darby *et al.*, 2005). Both loci pre-existed on the chromosome of *Y. pseudotuberculosis*, and their presence was most likely an essential prerequisite for the emergence of a bacterium transmitted by fleas. The acquisition of pFra in a *gmhA-hms*-positive chromosomal background thus allowed the recombinant *Y. pseudotuberculosis* progenitor to survive, to multiply and form a biofilm in the insect gut, and consequently to become a bacterium transmitted by fleas.

However, once injected by the flea into the dermis of a new host, this newly emerged organism had to deal with an environment drastically

different from that of the gut lumen. The second and crucial step in the emergence of *Y. pestis* has certainly been the acquisition of the pPla plasmid, which allowed the bacteria to cause bubonic plague instead of a primary septicemic disease of limited transmissibility. Acquisition of pPla by horizontal transfer enabled the bubonic form of disease and thus increased *Y. pestis* potential for epidemic spread (Sebbane *et al.*, 2006).

Therefore, the emergence and persistence of a flea-borne transmitted pathogen has probably resulted from the exceptional conjunction of different factors:

1. Co-infection of the intestine of a mammal with a bacterium that carried the conjugative pFra ancestor and with *Y. pseudotuberculosis*,
2. Transfer of this plasmid to *Y. pseudotuberculosis*,
3. Pre-existence in the recipient's genome of loci essential for flea blockage,
4. Dissemination of the recombinant *Y. pseudotuberculosis* (pFra) organism to the bloodstream of the infected animal,
5. Ingestion of the recombinant bacterium present in the blood by fleas living on the infected animal,
6. Subsequent transmission of the new microorganism by fleas,
7. Secondary acquisition of a plasmid (pPla) conferring an increased potential for epidemic spread.

8.4.2. Emergence of a Bacterium with an Exceptional Pathogenicity

The exceptional pathogenicity of *Y. pestis* and the particular type of disease it causes (the plague) may be explained by its new mode of transmission by flea bites (Carniel, 2003; Hinnebusch, 2004).

8.4.2.1. Pathogenesis

Y. pseudotuberculosis has a marked tropism for lymphoid tissues. When the bacterium penetrates the digestive tract, it translocates to the lymph node draining the intestine and multiplies there. This infection is called a mesenteric lymphadenitis and is clinically characterized by intestinal symptoms (abdominal pain, fever, and sometimes diarrhea). *Y. pestis* has inherited from its ancestor the same tropism for lymph nodes. Since this bacillus is injected into the dermis, it also reaches the draining lymph node (which has different locations depending on the site of the flea bite) and multiplies there. This infected lymph node is called a bubo and the disease is recognized as bubonic plague. Therefore, the mode of

penetration of the bacillus into the host organism determines the type of disease.

8.4.2.2. Severity of Infection

For an intestinal pathogen such as *Y. pseudotuberculosis*, the most efficient way to be transmitted to new hosts is to cause a diarrhea, which allows an efficient spread of the organisms in the environment and the secondary contamination of animals that consume the infected greenery. In contrast, since *Y. pestis* has a narrow flea-rodent-flea cycle, the only means for the species to be transmitted (and therefore to persist) is to spread to the blood of its host in order to be ingested by fleas. However, transmission of *Y. pestis* by fleas is not a very efficient process compared to the transmission of most other vector-borne pathogens (Lorange *et al.*, 2005). Furthermore, *Y. pestis* sepsis is rapidly fatal and the period during which fleas may become infected is brief. A high threshold bacteremia level thus needed to be attained to allow a complete transmission cycle. The new life cycle of *Y. pestis* required the bacterium to cause a potent septicemia. The arthropod-borne transmission of *Y. pestis* thus exerted a very strong selective pressure to exacerbate the pathogenic potential of this bacterium.

8.4.2.3. Mechanisms of Pathogenicity

The mechanisms underlying the capacity of *Y. pestis* to cause a fulminant septicemia are still unknown. The two additional *Y. pestis*-specific plasmids, pFra and pPla, may play some role but are not sufficient to explain the exceptional pathogenicity of the plague bacillus. Additional chromosomally encoded factors are most likely involved, one of them being the horizontally acquired filamentous phage YpfΦ. The analysis of the few genomic features that differentiate *Y. pestis* and *Y. pseudotuberculosis* should provide further clues for the identification of species-specific chromosomal regions that may participate in the exceptional pathogenicity of *Y. pestis*.

8.5. CONCLUSIONS

The sequential acquisition of mobile genetic elements by horizontal transfer has been of key importance for the gradual increase of virulence in *Yersinia*. The acquisition of the pYV plasmid has transformed a non-pathogenic environmental bacterium into an enteropathogen causing diseases of moderate severity. The subsequent acquisition of the HPI has conferred the capacity to spread systemically and to cause severe infections. In this genetic background, the horizontal acquisition of two plasmids and

a filamentous phage have led to the emergence of a flea-borne transmitted microorganism, known to be one of the most pathogenic bacteria for humans.

REFERENCES

Achtman, M., Zurth, K., Morelli, C., *et al.* (1999). *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proc Natl Acad Sci USA*, **96**, 14043–8.

Achtman, M., Morelli, G., Zhu, P., *et al.* (2004). Microevolution and history of the plague bacillus, *Yersinia pestis*. *Proc Natl Acad Sci USA*, **101**, 17837–42.

Aepfelbacher, M. (2004). Modulation of Rho GTPases by type III secretion system translocated effectors of *Yersinia*. *Rev Physiol Biochem Pharmacol*, **152**, 65–77.

Aili, M., Hallberg, B., Wolf-Watz, H., and Rosqvist, R. (2002). GAP activity of *Yersinia* YopE. *Methods Enzymol*, **359**, 359–70.

Antonenka, U., Nolting, C., Heesemann, J., and Rakin, A. (2005). Horizontal transfer of *Yersinia* high-pathogenicity island by the conjugative RP4 *attB* target-presenting shuttle plasmid. *Mol Microbiol*, **57**, 727–34.

Bercovier, H., Mollaret, H. H., Alonso, J. M., *et al.* (1980). Intra- and interspecies relatedness of *Yersinia pestis* by DNA hybridization and its relationship to *Yersinia pseudotuberculosis*. *Curr Microbiol*, **4**, 225–9.

Carniel, E. (2003). Evolution of pathogenic *Yersinia*, some lights in the dark. *Adv Exp Med Biol*, **529**, 3–12.

Chain, P. S., Carniel, E., Larimer, F. W., *et al.* (2004). Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. *Proc Natl Acad Sci USA*, **101**, 13826–31.

Chain, P. S. G., Hu, P., Malfatti, S. A., *et al.* (2006). Complete genome sequence of *Yersinia pestis* strains Antiqua and Nepal516: Evidence of gene reduction in an emerging pathogen. *J Bacteriol*, **188**, 4453–63.

Collyn, F., Marceau, M., and Simonet, M. (2004). YAPI, a new Pathogenicity Island in enteropathogenic yersiniae. In Carniel, E., and Hinnebusch, B. J. (Eds.) *Yersinia molecular and cellular biology*. Horizon Bioscience.

Collyn, F., Fukushima, H., Carnoy, C., Simonet, M., and Vincent, P. (2005). Linkage of the horizontally acquired *ypm* and *pil* genes in *Yersinia pseudotuberculosis*. *Infect Immun*, **73**, 2556–8.

Collyn, F., Guy, L., Marceau, M., Simonet, M., and Roten, C. A. (2006). Describing ancient horizontal gene transfers at the nucleotide and gene levels by comparative pathogenicity island genometrics. *Bioinformatics*, **22**, 1072–9.

Darby, C., Ananth, S. L., Tan, L., and Hinnebusch, B. J. (2005). Identification of *gmhA*, a *Yersinia pestis* gene required for flea blockage, by using a *Caenorhabditis elegans* biofilm system. *Infect Immun*, **73**, 7236–42.

Deng, W., Burland, V., Plunkett, G., *et al.* (2002). Genome sequence of *Yersinia pestis* KIM. *J Bacteriol*, **184**, 4601–11.

Derbise, A., Chenal-Francisque, V., Pouillot, F., *et al.* (2007). A horizontally acquired filamentous phage contributes to the pathogenicity of the plague bacillus. *Mol Microbiol*, **63**, 1145–57.

Drancourt, M., Roux, W., Dang, L. V., *et al.* (2004). Genotyping, orientalis-like *Yersinia pestis*, and plague pandemics. *Emerg Infect Dis*, **10**, 1585–92.

Fallman, M., and Gustavsson, A. (2005). Cellular mechanisms of bacterial internalization counteracted by *Yersinia*. *Int Rev Cytol*, **246**, 135–88.

Foultier, B., Troisfontaines, P., Muller, S., Opperdoes, F. R., and Cornelis, G. R. (2002). Characterization of the ysa pathogenicity locus in the chromosome of *Yersinia enterocolitica* and phylogeny analysis of type III secretion systems. *J Mol Evol*, **55**, 37–51.

Golubov, A., Neubauer, H., Nolting, C., Heesemann, J., and Rakin, A. (2004). Structural organization of the pFra virulence-associated plasmid of rhamnose-positive *Yersinia pestis*. *Infect Immun*, **72**, 5613–21.

Hacker, J., Blum-Oehler, G., Mühldorfer, I., and Tschäpe, H. (1997). Pathogenicity islands of virulent bacteria: structure, function, and impact on microbial evolution. *Mol Microbiol*, **23**, 1089–97.

Heesemann, J., Sing, A., and Trülzsch, K. (2006). *Yersinia*'s stratagem: targeting innate and adaptive immune defense. *Curr Opin Microbiol*, **9**, 55–61.

Hinnebusch, B. J. (2004). The evolution of flea-borne transmission in *Yersinia pestis*. In Carniel, E., and Hinnebusch, B. J. (Eds.). *Yersinia molecular and cellular biology*. Horizon Bioscience.

Hinnebusch, B. J., Perry, R. D., and Schwan, T. G. (1996). Role of the *Yersinia pestis* hemin storage (hms) locus in the transmission of plague by fleas. *Science*, **273**, 367–70.

Hu, P., Elliott, J., McCready, P., *et al.* (1998). Structural organization of virulence-associated plasmids of *Yersinia pestis*. *J Bacteriol*, **180**, 5192–202.

Korhonen, T. K., Kukkonen, M., Virkola, R., *et al.* (2004). The plasminogen activator Pla of *Yersinia pestis*: localized proteolysis and systemic spread. In Carniel, E., and Hinnebusch, B. J. (Eds.). *Yersinia molecular and cellular biology*. Horizon Bioscience.

Kutyrev, V., Mehigh, R. J., Motin, V. L., *et al.* (1999). Expression of the plague plasminogen activator in *Yersinia pseudotuberculosis* and *Escherichia coli*. *Infect Immun*, **67**, 1359–67.

Lesic, B., and Carniel, E. (2004). The High-Pathogenicity Island: a broad-host-range pathogenicity island. In Carniel, E., and Hinnebusch, B. J. (Eds.). *Yersinia molecular and cellular biology*. Horizon Bioscience.

Lesic, B., and Carniel, E. (2005). Horizontal transfer of the high-pathogenicity island of *Yersinia pseudotuberculosis*. *J Bacteriol*, **187**, 3352–8.

Lindler, L. E. (2004). The *Yersinia pestis*-specific plasmids pFra and pPla. In Carniel, E., and Hinnebusch, B. J. (Eds.) *Yersinia molecular and cellular biology*. Horizon Bioscience.

Lorange, E. A., Race, B. L., Sebbane, F., and Hinnebusch, B. J. (2005). Poor vector competence of fleas and the evolution of hypervirulence in *Yersinia pestis*. *J Infect Dis*, **191**, 1907–12.

MacIntyre, S., Knight, S. D., and Fooks, L. J. (2004). Structure, assembly and application of the polymeric F1 antigen of *Yersinia pestis*. In Carniel, E., and Hinnebusch, B. J. (Eds.). *Yersinia molecular and cellular biology*. Horizon Bioscience.

Marenne, M.-N., Mota, L. J., and Cornelis, G. R. (2004). The pYV plasmid and the Ysc-Yop type III secretion system. In Carniel, E., and Hinnebusch, B. J. (Eds.). *Yersinia molecular and cellular biology*. Horizon Bioscience.

Navarro, L., Alto, N. M., and Dixon, J. E. (2005). Functions of the *Yersinia* effector proteins in inhibiting host immune responses. *Curr Opin Microbiol*, **8**, 21–7.

Parkhill, J., Wren, B. W., Thomson, N. R., *et al.* (2001). Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature*, **413**, 523–7.

Perry, R. D., Straley, S. C., Fetherston, J. D., *et al.* (1998). DNA sequencing and analysis of the low-Ca^{2+}-response plasmid pCD1 of *Yersinia pestis* KIM5. *Infect Immun*, **66**, 4611–23.

Pouillot, F., Derbise, A., Kukkonen, M., *et al.* (2005). Evaluation of O-antigen inactivation on Pla activity and virulence of *Yersinia pseudotuberculosis* harbouring the pPla plasmid. *Microbiology*, **151**, 3759–68.

Prentice, M. B., James, K. D., Parkhill, J., *et al.* (2001). *Yersinia pestis* pFra shows biovar-specific differences and recent common ancestry with a *Salmonella enterica* serovar typhi plasmid. *J Bacteriol*, **183**, 2586–94.

Ramamurthi, K. S., and Schneewind, O. (2002). Type III protein secretion in *Yersinia* species. *Ann Rev Cell Dev Biol*, **18**, 107–33.

Schubert, S., Dufke, S., Sorsa, J., and Heesemann, J. (2004a). A novel integrative and conjugative element (ICE) of *Escherichia coli*: the putative progenitor of the *Yersinia* high-pathogenicity island. *Mol Microbiol*, **51**, 837–48.

Schubert, S., Rakin, A., and Heesemann, J. (2004b). The *Yersinia* high-pathogenicity island (HPI): evolutionary and functional aspects. *Int J Med Microbiol*, **294**, 83–94.

Sebbane, F., Jarrett, C. O., Gardner, D., Long, D., and Hinnebusch, B. J. (2006). Role of the *Yersinia pestis* plasminogen activator in the incidence of distinct septicemic and bubonic forms of flea-borne plague. *Proc Natl Acad Sci USA*, **103**, 5526–30.

Skurnik, M., Peippo, A., and Ervela, E. (2000). Characterization of the O-antigen gene clusters of *Yersinia pseudotuberculosis* and the cryptic O-antigen gene cluster of *Yersinia pestis* shows that the plague bacillus is most closely related to and has evolved from *Yersinia pseudotuberculosis* serotype O : 1b. *Mol Microbiol*, **37**, 316–30.

Sodeinde, O. A., Subrahmanyam, Y. V. B. K., Stark, K., *et al.* (1992). A surface protease and the invasive character of plague. *Science*, **258**, 1004–7.

Song, Y., Tong, Z., Wang, J., *et al.* (2004). Complete genome sequence of *Yersinia pestis* strain 91001, an isolate avirulent to humans. *DNA Res*, **11**, 179–97.

Viboud, G. I., and Bliska, J. B. (2005). *Yersinia* outer proteins: role in modulation of host cell signaling responses and pathogenesis. *Annu Rev Microbiol*, **59**, 69–89.

CHAPTER 9

Genomic or Pathogenicity Islands in *Streptococcus pneumoniae*

Barbara Albiger, Christel Blomberg, Jessica Dagerhamn, Staffan Normark, and Birgitta Henriques-Normark

9.1. INTRODUCTION

Bacterial genomes are no longer considered a stable and rigid DNA structure carrying the essential genetic information for survival, fitness, and transmission of the species. With the development of genomic high-throughput techniques such as sequencing of whole bacterial genomes (with a total in November 2007 of 597 complete microbial genomes that are sequenced and 879 genomes in progress; http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) and the bioinformatics tools, our view of bacterial genome plasticity has changed. Bacterial genomes are now regarded as highly flexible and dynamic structures that change in size, genetic content, and organization over time (Hanage *et al.*, 2006). This flexibility is also known as genome evolution (Groisman and Casadesus, 2005). It is thought that this evolution is driven by the adaptation of microorganisms to environmental niches, changes, or stresses and occurs via several mechanisms including point mutations, deletions, and gene acquisition/loss via horizontal gene transfer (Albiger *et al.*, 1999; Chen *et al.*, 2005). This latter process is associated with mobile genetic elements such as conjugative plasmids, bacteriophages, transposons, insertion sequence (IS) elements, and genomic islands (Albiger *et al.*, 1999; Frost *et al.*, 2005).

In Gram-negative bacteria, pathogenicity islands (PAI) are genomic regions that harbor one or more gene clusters encoding virulence-associated properties (reviewed in Gal-Mor and Finlay, 2006; Schmidt and Hensel, 2004). They are present in the genomes of pathogenic bacteria and usually absent from the same or closely related non-pathogenic species. Their sizes vary from a few kilobases (10 kb) to large chromosomal regions (up to 200 kb), and usually their base composition and codon usage (also referred

as G+C content) differ from the host genome. They are often integrated in proximity to tRNA or IS and are flanked by direct repeats (DR). Beside virulence genes, they frequently carry cryptic or functional motility genes such as integrases or transposases that can be involved in their integration or excision from the chromosome or plasmid. Because of the presence of motility genes, PAI are often unstable and form mosaic-like structures with several insertions occurring at different time points. These features apply to most PAI; however, some may lack one or more of them. PAI have primarily been found and described in Gram-negative bacteria, but have also been identified in Gram-positive organisms (reviewed in Schmidt and Hensel, 2004; Gal-Mor and Finlay, 2006). In this review we summarize the recently discovered PAI in the Gram-positive bacterium *S. pneumoniae* and discuss their contribution to bacterial colonization and virulence.

9.2. *STREPTOCOCCUS PNEUMONIAE*

9.2.1. Disease and Epidemiology

Streptococcus pneumoniae (also called the pneumococcus) is a Gram-positive diplococcus and a human-specific pathogen. It colonizes the human nasopharynx and causes diseases under certain conditions (Bogaert *et al.*, 2004). Among children, the carriage of *S. pneumoniae* is common and up to 70% of preschool children attending day-care centers may harbor these bacteria in their nasopharynx (Henriques-Normark *et al.*, 2003; Bogaert *et al.*, 2004). Pneumococci also account for 30 to 50% of otitis media cases (McCoy and Pettigrew, 2003; Hausdorff *et al.*, 2002), are a major cause of community-acquired pneumonia, and are responsible for more severe invasive infections also referred to as invasive pneumococcal disease (IPD), such as meningitis and pneumonia with or without bacteremia (Bogaert *et al.*, 2004). In fact, the World Health Organization (WHO) estimates that as many as 1 million children less than five years of age die worldwide each year as a result of pneumococcal diseases, especially in developing countries.

Pneumococci can be divided into at least 90 different serotypes based on the chemical structure of the polysaccharide capsule (CPS) (Bentley *et al.*, 2006). However, not only capsular serotype is important when studying genetic relatedness (clonality) between strains. For this purpose, sequence based methods have been developed such as multilocus sequence typing (MLST) (Enright and Spratt, 1999). MLST was proposed as a general approach for the molecular epidemiological investigation of bacterial pathogens to reflect their evolutionary and population biology (Hanage *et al.*, 2005; Enright and Spratt, 1998, 1999). This method is based on sequencing of seven

housekeeping genes in the pneumococcal genome. By this method, different families of strains are assigned an arbitrary sequence type (ST) reflecting their sequences in the seven loci. Thereby, clinical pneumococcal isolates may be divided into clones or clonal clusters, such as, genetically closely related bacteria will share the same or related ST (Enright and Spratt, 1998; Hanage *et al.*, 2005). However, strains belonging to the same clone or clonal cluster (i.e., differing at two or fewer of the seven loci) and thereby having the same or related ST can be of different serotypes because of capsular switches, where pneumococci acquire new capsular loci through transformation events (Coffey *et al.*, 1998a; Claverys *et al.*, 2000).

Some serotypes have been associated mainly with nasopharyngeal colonization (such as types 6A, 19F, and 35B) whereas others are more common in IPD (such as types 1, 4, 5, 7F, and 9V) (Sjostrom *et al.*, 2006; Sandgren *et al.*, 2004; Henriques *et al.*, 2000; Brueggemann *et al.*, 2003; Hausdorff *et al.*, 2005). However, it seems that not only capsular types are important for disease outcome but also other genetic properties associated with particular clonal types may play an important role for pathogenicity (Sandgren *et al.*, 2005). However, poorly invasive pneumococcal types that are good colonizers and therefore common among carriers, may still be important causes of pneumonia and invasive disease in humans because of their high abundance in the community (Sandgren *et al.*, 2004, 2005; Henriques-Normark *et al.*, 2003). Also, epidemiological studies reveal that some pneumococcal types are associated with invasive disease in previously healthy individuals (i.e., serotypes 1 and 7F), thereby acting as primary pathogens, while others primarily cause disease in compromised patients (Sjostrom *et al.*, 2006). Furthermore, some types are associated with more severe disease than others (Martens *et al.*, 2004; Henriques *et al.*, 2000). However, pneumococcal virulence as we measure it in animal mice models does not always correspond to the actual observations in human disease (Sandgren *et al.*, 2005).

9.2.2. Genomics

To date, only three pneumococcal genomes – TIGR4, an invasive strain of the serotype 4, D39, a serotype 2 strain and R6, a non-encapsulated derivative of the D39 strain – have been fully sequenced, been published, and are publicly available (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi; Hoskins *et al.*, 2001; Lanie *et al.*, 2007; Tettelin *et al.*, 2001). However, the genomes of sixteen additional pneumococcal strains of various serotypes (1, 3, 6B, 9V, 11, 14, 18, 19F, and 23F) are currently being sequenced (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). Both TIGR4 and D39 (the parental encapsulated strain of R6) are virulent in mice and generate an invasive disease

upon intranasal infection of several mice strains (Albiger *et al.*, 2005; Orihuela *et al.*, 2003, 2004a, 2004b; Sandgren *et al.*, 2005). Unfortunately, there is no complete genome information from a pneumococcal strain causing only carriage in mice (Sandgren *et al.*, 2005). Comparison between R6 and TIGR4 estimated the overall pneumococcal genome to be approximately 2.2 Mb in size with a G+C content of around 40% (Bruckner *et al.*, 2004; Hakenbeck *et al.*, 2001; Hoskins *et al.*, 2001; Tettelin *et al.*, 2001). It also showed that the R6 and TIGR4 genomes differ in size (Hoskins *et al.*, 2001; Tettelin *et al.*, 2001). While the R6 genome has 2,038,614 bp, the TIGR4 genome has 2,160,837 bp (Hoskins *et al.*, 2001; Tettelin *et al.*, 2001). They also diverge in their gene composition with a variability of 10% of their genes (Hakenbeck *et al.*, 2001; Bruckner *et al.*, 2004). Several comparative genome hybridization (CGH) microarray analyses of multiple pneumococcal strains confirmed that in any given strain there is a diversity of 8% to 10% of the genes when compared to the reference strain TIGR4 (Silva *et al.*, 2006; Bruckner *et al.*, 2004; Hakenbeck *et al.*, 2001). This diversity is believed to be the results of horizontal gene transfer associated with the abundant presence of IS, transposons, and repetitive elements (RUP and BOX) in the genome (Knutsen *et al.*, 2006; Oggioni and Claverys, 1999). Indeed, acquisition of both homologous DNA and DNA from closely related streptococci occurs by transformation since *S. pneumoniae* is naturally competent (Claverys and Havarstein, 2002; Claverys *et al.*, 2000; Hanage *et al.*, 2006). This genetic variability is evidenced by the expression of 90 distinct capsular serotypes and also by the emergence of antibiotic resistance (Bentley *et al.*, 2006; Henriques-Normark *et al.*, 2001, 2003; Claverys *et al.*, 2000). Most of the gene diversity is organized in large gene clusters, also called regions of diversity (RD), and up to 25 such RD have been identified by CGH when various clinical isolates were compared to the TIGR4 genome (Bruckner *et al.*, 2004; Hakenbeck *et al.*, 2001; Silva *et al.*, 2006). Since most of these RDs are often present only in few strains and are associated with atypical nucleotide content, they are the obvious candidates when searching for pathogenicity islands.

9.3. PUTATIVE PATHOGENICITY ISLANDS

9.3.1. Pneumococcal Pathogenicity Island-I (PPI-1) or RD6

The first suggested pneumococcal PAI (PPI-1) was identified a few months prior to the publication of the complete TIGR4 genome using a signature-tagged mutagenesis (STM) screen to identify new virulence-associated genes (Brown *et al.*, 2001). PPI-1 (SP1030-SP1063) was initially

described as a 27 kb pathogenicity island containing 31 genes including several virulence-associated genes, such as (1) an iron uptake locus *piaABCD* (SP1032-SP1035) and (2) a single gene SP1051 of unknown function, both of which are required for full virulence in a mouse model of pneumonia and systemic infection, as well as (3) a three-gene operon *phgABC* (SP1043-SP1045) of unknown function, required for growth in hyperosmotic medium and in vivo (Brown *et al.*, 2001, 2002, 2004a, 2004b; Brown and Holden, 2002). This PAI shares many features with the PAI in Gram-negative bacteria (see earlier description), including a different G+C content (32.6 ± 4%) as compared to the rest of the genome and the presence of motility genes (transposase and recombinase). Detailed analysis by polymerase chain reaction (PCR) and Southern blotting of 26 different strains of pneumococci (representing 12 different serotypes, 1, 2, 3, 4, 9N, 12, 14, 16, 17, 19F, and 23F) demonstrated that PPI-1 is a mosaic PAI with a variable region spanning between the open reading frames (ORFs) SP1046 and SP1064 (Brown *et al.*, 2004b). Within this variable region, either deletions or strain-specific insertions occur, suggesting that several gene rearrangements have been acquired, probably at different times. The region SP1047 and SP1063 is flanked by short 7-bp direct repeats (Brown *et al.*, 2004b). The three-gene operon *phgABC* (SP1043–SP1045) has a G+C content of 39.7%, which is close to the normal level for pneumococcal genomes, and is found to be present in *S. mitis* (Brown *et al.*, 2004a). It is therefore unlikely that this operon was acquired by a single event of horizontal gene transfer. It is more likely that *phgABC* is flanked by two regions that were horizontally acquired independently on two separate occasions (Brown *et al.*, 2004a). PPI-1 is therefore probably not one large PAI, but represents two independent smaller PAI (upstream and downstream) flanking a conserved region. However, the upstream PAI containing the iron uptake locus *piaABCD* (SP1032–SP1035) seems to be integrated in a stable manner into the chromosome since *piaABCD* is found in all the pneumococcal strains investigated, suggesting an essential function of this locus. However, mutation in *piaA* has only a marginal effect on growth in iron-depleted medium in vitro. Virulence in mice is also marginally affected because of the compensation by another ABC transporter, *piuBCDA*, involved in iron uptake located outside the PAI (Brown *et al.*, 2001, 2002; Brown and Holden, 2002).

9.3.2. The *rlrA* Islet or Pneumococcal Pathogenicity Island-II or RD4

Similarly to PPI-1, the *rlrA* islet was first identified through an STM screen that identified two new virulence genes *rlrA*, a rof-like regulator, and

srtD, a sortase (Hava and Camilli, 2002). Both these genes were shown to be required for colonization of the nasopharynx and for the development of pneumonia in mice (Hava and Camilli, 2002; Hava *et al.*, 2003b). The *rlrA* islet is a 14 kb PAI containing seven genes including *rlrA* and *srtD*. *rlrA* positively regulates the divergent transcription of six downstream genes, encoding three cell-wall anchored surface proteins with LPTXG motifs (RrgA, RrgB, and RrgC) and three sortase (SrtB, SrtC, and SrtD) (Hava *et al.*, 2003a; Hemsley *et al.*, 2003). The *rlrA* islet is negatively regulated by MgrA, a negative repressor located outside the islet (Hemsley *et al.*, 2003). The island is flanked by two direct copies of IS*1167* (Hava *et al.*, 2003a; Hemsley *et al.*, 2003). Recently, it was demonstrated that the *rlrA* islet encodes pilus-like structures on the pneumococcal surface (Barocchi *et al.*, 2006; LeMieux *et al.*, 2006). These pneumococcal pili have been shown to be important for adherence to mucosal cells and colonization of the mouse upper airways, but also to influence virulence in mice and to evoke a strong host inflammatory response upon intraperitoneal challenge with piliated bacteria (Barocchi *et al.*, 2006; Hava and Camilli, 2002). By immunogold electron microscopy, RrgB was shown to be the major pilin subunit that forms the core of the pilus and RrgA to be an ancillary protein that seems to be incorporated at regular interval into the pilus, while RrgC was located preferentially at the tip of the pilus (Barocchi *et al.*, 2006; LeMieux *et al.*, 2006). The precise role of these proteins, as well as the mechanisms of their incorporation into the pilus structure, is not yet fully understood and is under investigation. However, RrgA has been demonstrated to be incorporated into the pilus as a minor subunit in a *SrtD*-dependent manner (LeMieux *et al.*, 2006). Pilus structures and pilus islands with organizations similar to that of the *rlrA* islet were recently found in *Streptococcus agalactiae* (Group B *Streptococcus*, GBS), in *Streptococcus pyogenes* (Group A *Streptococcus*, GAS), and in *Enterococcus faecalis* (Gaspar and Ton-That, 2006; Lauer *et al.*, 2005; Mora *et al.*, 2005; Nallapareddy *et al.*, 2006; Rosini *et al.*, 2006; Dramsi *et al.*, 2006).

Bioinformatic analysis has found some evidence that genes encoding putative pilus-like structures can be found in other Gram-positive bacterial genomes, such as *S. suis* (our unpublished data), but the expression of pili in these bacteria is yet to be demonstrated (Telford *et al.*, 2006; Scott and Zahner, 2006). However, it was not found in the other closely related less pathogenic streptococcal species such as *Streptococcus gordonii*, *S. oralis*, and *S. mitis* (Telford *et al.*, 2006; Scott and Zahner, 2006). Analysis and comparison of the pilus island in all streptococcal species suggest that it has been acquired via horizontal gene transfer as an entity that carries all the

genes necessary for its biosynthesis and its expression (Telford *et al.*, 2006). However, the origin of this island remains unclear.

The *rlrA* islet is only present in a subset of pneumococcal serotypes and associated with particular clonal type (Telford *et al.*, 2006; Barocchi *et al.*, 2006). A correlation exists between some of these piliated clonal types and nasopharyngeal colonization and spread in the community (Sjostrom *et al.*, 2007).

9.4. REGIONS OF DIVERSITY AS PUTATIVE PATHOGENICITY ISLANDS

Thirteen RD are present in the TIGR4 genome and absent in R6 (Obert *et al.*, 2006; Tettelin *et al.*, 2001; Hakenbeck *et al.*, 2001; Bruckner *et al.*, 2004). They have been suggested to be putative PAIs and they are scattered on the genome (Figure 9.1). We now describe and discuss in detail some of these RDs for which functional data are available. Most of the other RDs encode putative ORFs of unknown function and have not been studied so far.

9.4.1. RD1

RD1 (SP0067-SP0074) is a small region containing only six genes including ZmpC (SP0071), a zinc metalloproteinase and CbpI (SP0069), a putative choline-binding protein of unknown function (Paterson *et al.*, 2006; Oggioni *et al.*, 2003). Pneumococci express up to four large zinc metalloproteinases on their surfaces: IgA protease, ZmpB, ZmpC, and ZmpD (Blue *et al.*, 2003; Oggioni *et al.*, 2003; Wani *et al.*, 1996). IgA protease cleaves human IgA1 in the hinge region and is important for the development of pneumonia and sepsis as well as for in vitro adherence to epithelial cells (Wani *et al.*, 1996; Kilian *et al.*, 1996). The substrates for both ZmpB and ZmpD have not yet been identified; however, *zmpB*-deficient mutants are attenuated in murine models of pneumonia and sepsis (Blue *et al.*, 2003). ZmpC is a metalloproteinase that cleaves human matrix metalloproteinase 9 (MMP-9), which itself cleaves extracellular matrix gelatin and collagen (Blue *et al.*, 2003; Camilli *et al.*, 2006; Oggioni *et al.*, 2003; Chiavolini *et al.*, 2003). MMP-9 is activated by proteinase cleavage and has been suggested to play a role in opening the blood-brain barrier during inflammation and tissue invasion by tumor cells. It was initially thought that ZmpC could play a role in mediating pneumococcal invasion into tissue. However, although ZmpC has been shown to contribute to virulence in an experimental mouse model of pneumonia and sepsis, it was less virulent as compared to the two other

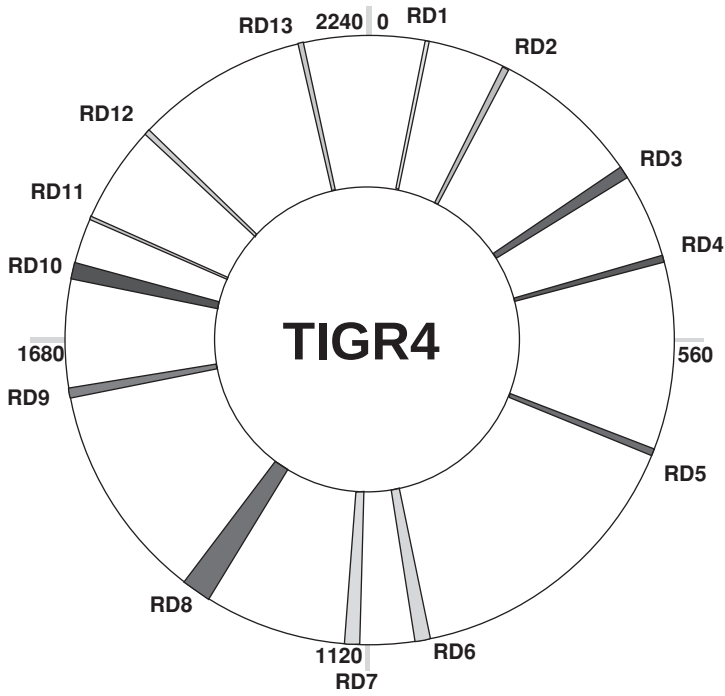


Figure 9.1. Circular representation of the *Streptococcus pneumoniae* TIGR4 genome. The TIGR4 genome contains an estimated 2,236 open reading frames. Comparative genome hybridizations using microarrays have identified 13 major genomic regions of diversity (RD) between the TIGR4 and R6 strains. The 13 RD are represented on the outer circle. The RD discussed in this review are the following: (i) RD1 (SP0067–SP0074) contains a putative metalloproteinase ZmpC and a choline binding protein CbpI; (ii) RD3 (SP0343–SP0365) encodes for the capsular polysaccharide biosynthesis proteins of serotype 4; (iii) RD4 (SP0460–SP0469) is the pilus encoding islet or the *rlrA* islet; (iv) RD6 or the pneumococcal pathogenicity island-I spans from SP1032 to SP1067 and probably represents two independent smaller PAI (upstream and downstream) flanking the conserved region *phgABC* (SP1043-SP1045); (v) RD8 (SP1314–SP1352) harbors a putative neuraminidase, NanC; and (vi) RD10 (SP1753–SP1772) contains an operon homologous to the *gspB-secY2A2* from *S. gordonii*.

pneumococcal metalloproteinases, IgA protease and ZmpB, suggesting that there may be redundancy of function resulting in compensation (Chiavolini *et al.*, 2003).

Investigation of the respective distribution of *igA*, *zmpB*, and *zmpC* showed that while *igA* and *zmpB* are present in all pneumococcal strains, *zmpC* is found only in 18–25% strains (Camilli *et al.*, 2006). The ZmpC was

mainly found in two serotypes, 8 and 11A, belonging to the same clonal cluster (Camilli *et al.*, 2006). No correlation has been found so far between the presence/absence of *zmpC* and whether the strains were isolated from meningitis, pneumonia, or sepsis or from healthy carriers.

9.4.2. Capsular Operon (RD3)

The capsule polysaccharide is regarded as the major virulence factor of pneumococci, and non-encapsulated strains are considered avirulent, although they may be found in humans. The capsule protects the bacteria from phagocytosis by polymorphonuclear leucocytes and as many as 90 different polysaccharide capsular types have been described so far (Bentley *et al.*, 2006). Genes encoding for biosynthesis of the capsule are organized in cassettes (*cps* locus), which are always inserted into the same chromosomal location between the genes *dexB* and *aliA* (with the exception of serotype 37). The *cps* locus varies in size from 10 kb to 30 kb (Bentley *et al.*, 2006). Recently all *cps* loci of all 90 serotypes were sequenced (Bentley *et al.*, 2006). It is the first time that a complete repertoire of capsular biosynthetic genes is available for a pathogen. The analysis showed that in all cases, the *cps* locus is composed of a region of low G+C content defining the serotype-specific feature of each locus and is always flanked by intact or disrupted mobile element such as IS. The low percentage of G+C content suggests that the different capsular loci have been generated by the introduction of novel *cps* genes into the chromosome by horizontal gene transfer from other species. Furthermore, capsular exchange between two pneumococcal strains, also called serotype switching, is not uncommon, and it is believed to be mediated by recombinational replacements within the capsular biosynthesis (*cps*) operon and its flanking regions following natural transformation by DNA from other pneumococcal strains (Coffey *et al.*, 1998a; Claverys *et al.*, 2000). For example, among the internationally distributed antibiotic resistant clones of *S. pneumoniae*, Coffey *et al.* found serotype 19F variants of the major multi-resistant Spain^{23F} clones that had emerged at multiple occasions rather than arising from a single ancestral serotype 19F variant (Coffey *et al.*, 1999, 1998b). Also, Sandgren *et al.* (2004) found such capsular switches between serotype 19F and 9V isolates with ST156 and ST162 in a large epidemiological study in Sweden. Capsular switches allow the serotype variants to avoid opsonization and neutralization by antibodies against serotypes to which the host was previously exposed. This has strong implications for the efficiency of current vaccines based on a limited number of capsular polysaccharides as well as for the development of new vaccines (Tai, 2006). Different studies

have recently highlighted the emergence of non-vaccine serotypes (NVT), so-called replacement after the introduction of the seven-valent conjugated polysaccharide vaccines (Kyaw *et al.*, 2006; Beall *et al.*, 2006; Pebody *et al.*, 2005). More studies are needed to evaluate the risk for replacement with new serotypes resulting in treatment failures in the vaccines era.

9.4.3. RD8

RD8 has a size of approximately 40 kb, is flanked by remnant IS elements, and contains a putative transposase and 31 predicted ORFs, including a putative neuraminidase called NanC (Pettigrew *et al.*, 2006). Two pneumococcal neuraminidases have been previously described for pneumococci, NanA and NanB, both located outside RD8 (Manco *et al.*, 2006). Both neuraminidases are thought to cleave sialic acid-containing molecules on the host cell surface and thereby either mediate pneumococcal adherence by stripping the host cell surface to expose potential receptors or disarm potential antibacterial proteins such as lactoferrin or immunoglobulin A2 (Manco *et al.*, 2006). Recently it was shown that both NanA and NanB are important for persistent colonization of the upper and lower respiratory tract of mice as well as for survival in blood (Manco *et al.*, 2006). Furthermore, NanA has been shown to remove sialylated moieties from lipopolysaccharides of *Haemophilus influenzae* and *Neisseria meningitidis*, two Gram-negative bacteria that inhabit the same ecological niche as pneumococci, making them more susceptible to complement-mediated clearance (Shakhnovich *et al.*, 2002; King *et al.*, 2006). A study investigating the distribution of pneumococcal neuraminidase in relation to ST and serotype showed that *nanC* is only found in 51% of the strains analyzed while *nanA* and *nanB* are found in 100% and 96% respectively (Pettigrew *et al.*, 2006). Even though the exact role of NanC in vivo is not known to date, a higher prevalence of *nanC* was found in invasive isolates as compared to carriage isolates, especially in cerebrospinal fluid (CSF) isolates, suggesting that NanC might be required to cause meningitis (Pettigrew *et al.*, 2006).

Using CGH-based analysis of 42 invasive and 30 noninvasive clinical isolates of serotype 6A, 6B, and 14, Obert *et al.* (2006) proposed that RD8 is composed of two RD located precisely next to each other in the TIGR4 genome, RD8a (SP1315–SP1331) and RD8b (SP1332–SP1351). RD8a correlated to the invasive phenotype and RD8b correlated to the noninvasive phenotype. RD8a was therefore suggested to be a PAI. Hence, it is therefore interesting that *nanC* is located in RD8a and was mainly found by Obert *et al.* among isolates of serotype 6B and 14 where more than half of the invasive

isolates of serotype 14 (8 out of 12 isolates) analyzed were from CSF (Obert *et al.*, 2006). This suggests that NanC may indeed contribute to the capacity of pneumococci to cause meningitis. However, another CGH analysis, using three invasive isolates of serotype 14 belonging to the same clone (ST124), did not come to the same conclusions (Silva *et al.*, 2006). Here the authors observed that RD8a was absent in two of the three invasive isolates investigated. However, the isolates were from blood and not from CSF. Hence, whether or not NanC plays an important role in meningitis remains to be shown.

9.4.4. RD10

RD10 is a 36 kb region encoding seventeen putative ORFs of which three were identified by STM (Hava and Camilli, 2002). Bioinformatic analysis showed that RD10 is homologous to an operon in *S. gordonii*, the *gspB-secY2A2* operon (Obert *et al.*, 2006; Takamatsu *et al.*, 2004, 2005a, 2005b, 2006; Bensing *et al.*, 2004). This operon encodes GspB (also known as Hsa), a serine-rich surface glycoprotein that mediates the binding of this organism to platelet membrane glycoprotein GP Iba via sialic acid moieties (Bensing *et al.*, 2004; Takamatsu *et al.*, 2004, 2005a, 2005b, 2006). The *gspB* gene in *S. gordonii* is 9.2 kb and is adjacent to a region necessary for surface expression of the mature glycoprotein, the *secY2A2* locus, which encodes eleven proteins involved either in glycosylation of GspB or in the secretion of GspB (Takamatsu *et al.*, 2004, 2005a, 2005b, 2006; Bensing *et al.*, 2004). TIGR4 possesses a GspB homologue called PsrP (for pneumococcal serine-rich repeat protein) (Obert *et al.*, 2006).

The organization of GspB and its homologues is highly conserved among all species in which this operon has been identified, and PsrP is no exception: (1) a large N-terminal signal peptide of 72 amino acids (aa), (2) two serine-rich regions (SRR1 of 49 aa and SRR2 of 4,319 aa), (3) an intermediate region rich in either basic or acidic amino acids residues between SRR1 and SRR2 called a basic region (BR, of 272 aa), and (4) a C-terminal cell wall anchoring domain (62 aa). As in *S. gordonii*, the region adjacent of PsrP encodes for genes that are likely to be involved in the glycosylation and the export of the protein (Takamatsu *et al.*, 2004, 2005a, 2005b, 2006; Bensing *et al.*, 2004). However, RD10 in TIGR4 encodes for seven additional glycosyl transferases (Obert *et al.*, 2006). The role of these extra enzymes is not yet known. *S. gordonii* and related species of viridans group streptococci are part of the commensal flora of the oral cavity and play significant roles in the formation of dental plaques (Burne, 1998). Furthermore, the GspB/Hsa protein

family has also been shown to mediate bacterial adherence to O-glycosylated mucin-type sialoglycoproteins, such as salivary mucin and leukosialin, the major surface glycoprotein of human polymorphonuclear leukocytes (Takamatsu *et al.*, 2006; Jakubovics *et al.*, 2005). Interestingly, PsrP has been shown to be required for virulence in lung infection but not for nasopharyngeal colonization in a mouse model of intranasal infection, suggesting that PsrP may have different affinities for other glycoproteins as compared to GspB such as glycoproteins expressed in the lower respiratory tract rather than in the nasopharynx (Obert *et al.*, 2006). We can only speculate that the additional glycosyltransferases may be responsible for this differential affinity, but that remains to be demonstrated. Furthermore, *S. gordonii* is also a leading cause of infective endocarditis, and the binding of this organism to human platelets is thought to be a major virulence factor in the pathogenesis of this disease (Herzberg *et al.*, 1997). Endocarditis due to pneumococci occurs, albeit at a low frequency (Kan *et al.*, 2006). It would be interesting to investigate the presence of this operon in isolates collected from patients with endocarditis.

9.5. CONCLUDING REMARKS

The correlation between invasive disease potential and pneumococcal serotype has been extensively studied, unlike the correlation between disease potential and the presence of PAI and/or virulence-associated genes (Sandgren *et al.*, 2004; Hausdorff, 2002; Hausdorff *et al.*, 2001, 2002, 2005; Brueggemann *et al.*, 2003, 2004; Brueggemann and Spratt, 2003; Sjostrom *et al.*, 2006). In the era of new and powerful tools such as genome sequencing and profiling, comparative genome analysis, and molecular epidemiological typing, we hope to identify genetic determinants that will explain why certain pneumococcal strains cause disease and others only carriage. We are, however, far from understanding the relative contribution of each of the putative PAIs in human invasive pneumococcal disease.

We know that the polysaccharide capsule is important for disease outcome, but recent data suggest that also the genetic makeup of the bacteria is a major contributor to virulence (Sandgren *et al.*, 2005; Silva *et al.*, 2006). The serotype distribution varies with geographic area and time period studied, as well as with age and pneumococcal disease (Hausdorff *et al.*, 2001; Brueggemann and Spratt, 2003; Brueggemann *et al.*, 2004; Henriques-Normark *et al.*, 2001, 2003). Some serotypes and clones are more common in invasive disease, while others mainly lead to carriage (Sandgren *et al.*, 2004, 2005; Brueggemann *et al.*, 2003; Meats *et al.*, 2003). Furthermore, isolates belonging to invasive pneumococcal serotypes such as types

1 and 7F are more closely genetically related as compared to other serotypes such as type 19F, found to be more prone to cause carriage (Sandgren *et al.*, 2004; Sjostrom *et al.*, 2006). Similarly, serogroups 3, 6, 14, 18, 19, and 23 are most common among cases of otitis media, accounting for over 73% of middle ear isolates (McCoy and Pettigrew, 2003; Hausdorff *et al.*, 2002). Recent studies employing CGH techniques have revealed that the genetic diversity found within serotypes and clones is probably higher than it has been previously estimated, leading to divergent results between research groups analyzing different collections of pneumococcal isolates (Silva *et al.*, 2006; Pettigrew and Fennie, 2005; Orihuela *et al.*, 2004b; Obert *et al.*, 2006; Hakenbeck *et al.*, 2001; Bruckner *et al.*, 2004). To better understand the pathogenesis of pneumococcal disease, more studies on the molecular epidemiology of pneumococcal infections in relation to disease and virulence are needed; isolates with different characteristics should be compared using technologies with high discriminatory power such as CGH.

Interestingly, most suggested or potential PAI identified in pneumococci so far harbor genes that are required for full virulence in animal models. However, they are not present in all pneumococcal strains. Also, isolates belonging to certain serotypes and clones mainly causing carriage may cause disease in susceptible individuals, showing that host factors are also important for disease outcome (Albiger *et al.*, 2005, 2006). However, the presence and the conservation of specific genes among strains belonging to the same serotype and/or the same clone may be driven by an evolutionary pressure to adapt to environmental conditions or niches such as the nasopharynx, which may modulate their fitness only in this specific habitat. The pilus islet, for example, has mainly been found in isolates belonging to serotypes commonly found among carriers, whereas serotypes with a high invasive disease potential such as clones of type 1 and 7F lack the islet (Sjostrom *et al.* 2007).

Apart from the capsule operon, no PAI or suggested PAI have been strictly correlated to invasive pneumococcal disease potential in humans. However, most of the other RD not discussed in this review encode putative ORFs of unknown function and have not yet been studied. We therefore cannot exclude that these RD are important for pneumococcal virulence. Future studies will reveal to what extent different combinations of pneumococcal genomic islands contribute to invasive disease. Also, we await additional genome information from highly invasive strains such as isolates of serotype 1, showing ability to cause invasive disease in previously healthy individuals as well as to cause epidemic outbreaks (Sjostrom *et al.*, 2006). Finally, we expect to find additional PAI when more pneumococcal isolates are being sequenced.

REFERENCES

Albiger B., Hubert J. C., and Lett M. C. (1999). Identification of the plasmid-mobilization potential of the strain *Klebsiella pneumoniae ozenae* KIIIA isolated from a polluted aquatic environment. *Plasmid*, **41**, 30–9.

Albiger, B., Sandgren, A., Katsuragi, H., *et al.* (2005). Myeloid differentiation factor 88-dependent signalling controls bacterial growth during colonization and systemic pneumococcal disease in mice. *Cell Microbiol*, **7**, 1603–15.

Albiger, B., Dahlberg, S., Sandgren, A., *et al.* (2006). Toll-like receptor 9 acts at an early stage in host defence against pneumococcal infection. *Cell Microbiol*, **9**, 633–44.

Barocchi, M. A., Ries, J., Zogaj, X., *et al.* (2006). A pneumococcal pilus influences virulence and host inflammatory responses. *Proc Natl Acad Sci USA*, **103**, 2857–62.

Beall, B., McEllistrem, M. C., Gertz, R. E., Jr., *et al.* (2006). Pre- and postvaccination clonal compositions of invasive pneumococcal serotypes for isolates collected in the United States in 1999, 2001, and 2002. *J Clin Microbiol*, **44**, 999–1017.

Bensing, B. A., Lopez, J. A., and Sullam, P. M. (2004). The *Streptococcus gordonii* surface proteins GspB and Hsa mediate binding to sialylated carbohydrate epitopes on the platelet membrane glycoprotein Iba. *Infect Immun*, **72**, 6528–37.

Bentley, S. D., Aanensen, D. M., Mavroidi, A., *et al.* (2006). Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. *PLoS Genet*, **2**, e31.

Blue, C. E., Paterson, G. K., Kerr, A. R., *et al.* (2003). ZmpB, a novel virulence factor of *Streptococcus pneumoniae* that induces tumor necrosis factor alpha production in the respiratory tract. *Infect Immun*, **71**, 4925–35.

Bogaert, D., De Groot, R., and Hermans, P. W. (2004). *Streptococcus pneumoniae* colonisation: the key to pneumococcal disease. *Lancet Infect Dis*, **4**, 144–54.

Brown, J. S., and Holden, D. W. (2002). Iron acquisition by Gram-positive bacterial pathogens. *Microb Infect*, **4**, 1149–56.

Brown, J. S., Gilliland, S. M., and Holden, D. W. (2001). A *Streptococcus pneumoniae* pathogenicity island encoding an ABC transporter involved in iron uptake and virulence. *Mol Microbiol*, **40**, 572–85.

Brown, J. S., Gilliland, S. M., Ruiz-Albert, J., and Holden, D. W. (2002). Characterization of *pit*, a *Streptococcus pneumoniae* iron uptake ABC transporter. *Infect Immun*, **70**, 4389–98.

Brown, J. S., Gilliland, S. M., Basavanna, S., *et al.* (2004a). *phgABC*, a three-gene operon required for growth of *Streptococcus pneumoniae* in hyperosmotic medium and in vivo. *Infect Immun*, **72**, 4579–88.

Brown, J. S., Gilliland, S. M., Spratt, B. G., and Holden, D. W. (2004b), A locus contained within a variable region of pneumococcal pathogenicity island 1 contributes to virulence in mice. *Infect Immun*, **72**, 1587–93.

Bruckner, R., Nuhn, M., Reichmann, P., Weber, B., and Hakenbeck, R. (2004) Mosaic genes and mosaic chromosomes-genomic variation in *Streptococcus pneumoniae*. *Int J Med Microbiol*, **294**, 157–68.

Brueggemann, A. B., and Spratt, B. G. (2003). Geographic distribution and clonal diversity of *Streptococcus pneumoniae* serotype 1 isolates. *J Clin Microbiol*, **41**, 4966–70.

Brueggemann, A. B., Griffiths, D. T., Meats, E., *et al.* (2003). Clonal relationships between invasive and carriage *Streptococcus pneumoniae* and serotype- and clone-specific differences in invasive disease potential. *J Infect Dis*, **187**, 1424–32.

Brueggemann, A. B., Peto, T. E., Crook, D. W., *et al.* (2004). Temporal and geographic stability of the serogroup-specific invasive disease potential of *Streptococcus pneumoniae* in children. *J Infect Dis*, **190**, 1203–11.

Burne, R. A. (1998). Oral streptococci . . . products of their environment. *J Dent Res*, **77**, 445–52.

Camilli, R., Pettini, E., Grosso, M. D., *et al.* (2006). Zinc metalloproteinase genes in clinical isolates of *Streptococcus pneumoniae*: association of the full array with a clonal cluster comprising serotypes 8 and 11A. *Microbiology*, **152**, 313–21.

Chen, I., Christie, P. J., and Dubnau, D. (2005). The ins and outs of DNA transfer in bacteria. *Science*, **310**, 1456–60.

Chiavolini, D., Memmi, G., Maggi, T., *et al.* (2003). The three extra-cellular zinc metalloproteinases of *Streptococcus pneumoniae* have a different impact on virulence in mice. *BMC Microbiol*, **3**, 14.

Claverys, J. P., and Havarstein, L. S. (2002). Extracellular-peptide control of competence for genetic transformation in *Streptococcus pneumoniae*. *Front Biosci*, **7**, d1798–814.

Claverys, J. P., Prudhomme, M., Mortier-Barriere, I., and Martin, B. (2000). Adaptation to the environment: *Streptococcus pneumoniae*, a paradigm for recombination-mediated genetic plasticity? *Mol Microbiol*, **35**, 251–9.

Coffey, T. J., Enright, M. C., Daniels, M., *et al.* (1998a). Recombinational exchanges at the capsular polysaccharide biosynthetic locus lead to frequent serotype changes among natural isolates of *Streptococcus pneumoniae*. *Mol Microbiol*, **27**, 73–83.

Coffey, T. J., Enright, M. C., Daniels, M., *et al.* (1998b). Serotype 19A variants of the Spanish serotype 23F multiresistant clone of *Streptococcus pneumoniae*. *Microb Drug Resist*, **4**, 51–5.

Coffey, T. J., Daniels, M., Enright, M. C., and Spratt, B. G. (1999). Serotype 14 variants of the Spanish penicillin-resistant serotype 9V clone of *Streptococcus pneumoniae* arose by large recombinational replacements of the *cpsA-pbp*1a region. *Microbiology*, **145**, 2023–31.

Dramsi, S., Caliot, E., Bonne, I., *et al*. (2006). Assembly and role of pili in group B streptococci. *Mol Microbiol*, **60**, 1401–13.

Enright, M. C., and Spratt, B. G. (1998). A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. *Microbiology*, **144**, 3049–60.

Enright, M. C., and Spratt, B. G. (1999). Multilocus sequence typing. *Trends Microbiol*, **7**, 482–7.

Frost, L. S., Leplae, R., Summers, A. O., and Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol*, **3**, 722–32.

Gal-Mor, O., and Finlay, B. B. (2006). Pathogenicity islands: a molecular toolbox for bacterial virulence. *Cell Microbiol*, **8**, 1707–19.

Gaspar, A. H., and Ton-That, H. (2006). Assembly of distinct pilus structures on the surface of *Corynebacterium diphtheriae*. *J Bacteriol*, **188**, 1526–33.

Groisman, E. A., and Casadesus, J. (2005). The origin and evolution of human pathogens. *Mol Microbiol*, **56**, 1–7.

Hakenbeck, R., Balmelle, N., Weber, B., *et al*. (2001). Mosaic genes and mosaic chromosomes: intra- and interspecies genomic variation of *Streptococcus pneumoniae*. *Infect Immun*, **69**, 2477–86.

Hanage, W. P., Kaijalainen, T., Herva, E., *et al*. (2005). Using multilocus sequence data to define the pneumococcus. *J Bacteriol*, **187**, 6223–30.

Hanage, W. P., Fraser, C., and Spratt, B. G. (2006). The impact of homologous recombination on the generation of diversity in bacteria. *J Theor Biol*, **239**, 210–9.

Hausdorff, W. P. (2002). Invasive pneumococcal disease in children: geographic and temporal variations in incidence and serotype distribution. *Eur J Pediatr*, **161**(Suppl 2), S135–9.

Hausdorff, W. P., Siber, G., and Paradiso, P. R. (2001). Geographical differences in invasive pneumococcal disease rates and serotype frequency in young children. *Lancet*, **357**, 950–2.

Hausdorff, W. P., Yothers, G., Dagan, R., *et al*. (2002). Multinational study of pneumococcal serotypes causing acute otitis media in children. *Pediatr Infect Dis J*, **21**, 1008–16.

Hausdorff, W. P., Feikin, D. R., and Klugman, K. P. (2005). Epidemiological differences among pneumococcal serotypes. *Lancet Infect Dis*, **5**, 83–93.

Hava, D. L., and Camilli, A. (2002). Large-scale identification of serotype 4 *Streptococcus pneumoniae* virulence factors. *Mol Microbiol*, **45**, 1389–406.

Hava, D. L., Hemsley, C. J., and Camilli, A. (2003a). Transcriptional regulation in the *Streptococcus pneumoniae rlrA* pathogenicity islet by RlrA. *J Bacteriol*, **185**, 413–21.

Hava, D. L., LeMieux, J., and Camilli, A. (2003b). From nose to lung: the regulation behind *Streptococcus pneumoniae* virulence factors. *Mol Microbiol*, **50**, 1103–10.

Hemsley, C., Joyce, E., Hava, D. L., Kawale, A., and Camilli, A. (2003). MgrA, an orthologue of Mga, acts as a transcriptional repressor of the genes within the *rlrA* pathogenicity islet in *Streptococcus pneumoniae*. *J Bacteriol*, **185**, 6640–7.

Henriques, B., Kalin, M., Ortqvist, A., *et al.* (2000). Molecular epidemiology of *Streptococcus pneumoniae* causing invasive disease in 5 countries. *J Infect Dis*, **182**, 833–9.

Henriques-Normark, B., Kalin, M., Ortqvist, A., *et al.* (2001). Dynamics of penicillin-susceptible clones in invasive pneumococcal disease. *J Infect Dis*, **184**, 861–9.

Henriques-Normark, B., Christensson, B., Sandgren, A., *et al.* (2003). Clonal analysis of *Streptococcus pneumoniae* nonsusceptible to penicillin at day-care centers with index cases, in a region with low incidence of resistance: emergence of an invasive type 35B clone among carriers. *Microb Drug Resist*, **9**, 337–44.

Herzberg, M. C., Meyer, M. W., Kilic, A., and Tao, L. (1997). Host-pathogen interactions in bacterial endocarditis: streptococcal virulence in the host. *Adv Dent Res*, **11**, 69–74.

Hoskins, J., Alborn, W. E. Jr., Arnold, J., *et al.* (2001). Genome of the bacterium *Streptococcus pneumoniae* strain R6. *J Bacteriol*, **183**, 5709–17.

Jakubovics, N. S., Kerrigan, S. W., Nobbs, A. H., *et al.* (2005). Functions of cell surface-anchored antigen I/II family and Hsa polypeptides in interactions of *Streptococcus gordonii* with host receptors. *Infect Immun*, **73**, 6629–38.

Kan, B., Ries, J., Normark, B. H., *et al.* (2006). Endocarditis and pericarditis complicating pneumococcal bacteraemia, with special reference to the adhesive abilities of pneumococci: results from a prospective study. *Clin Microbiol Infect*, **12**, 338–44.

Kilian, M., Reinholdt, J., Lomholt, H., Poulsen, K., and Frandsen, E. V. (1996) Biological significance of IgA1 proteases in bacterial colonization and pathogenesis: critical evaluation of experimental evidence. *APMIS*, **104**, 321–38.

King, S. J., Hippe, K. R., and Weiser, J. N. (2006). Deglycosylation of human glycoconjugates by the sequential activities of exoglycosidases expressed by *Streptococcus pneumoniae*. *Mol Microbiol*, **59**, 961–74.

Knutsen, E., Johnsborg, O., Quentin, Y. Claverys, J. P., and Havarstein, L. S. (2006). BOX-elements modulate gene expression in *Streptococcus*

pneumoniae: impact on the fine-tuning of competence development. *J Bacteriol*, **88**, 8307–12.

Kyaw, M. H., Lynfield, R., Schaffner, W., *et al.* (2006). Effect of introduction of the pneumococcal conjugate vaccine on drug-resistant *Streptococcus pneumoniae*. *N Engl J Med*, **354**, 1455–63.

Lanie J. A., Ng W. L., Kazmierczak K. M., Andrzejewski T. M., *et al.* (2007). Genome sequence of Avery's virulent serotype 2 strain D39 of *Streptococcus pneumoniae* and comparison with that of unencapsulated laboratory strain R6. *J Bacteriol*, **189**, 38-51.

Lauer, P., Rinaudo, C. D., Soriani, M., *et al.* (2005). Genome analysis reveals pili in Group B *Streptococcus*. *Science*, **309**, 105.

LeMieux, J., Hava, D. L., Basset, A., and Camilli, A. (2006). RrgA and RrgB are components of a multisubunit pilus encoded by the *Streptococcus pneumoniae rlrA* pathogenicity islet. *Infect Immun*, **74**, 2453–6.

Manco, S., Hernon, F., Yesilkaya, H., *et al.* (2006). Pneumococcal neuraminidases A and B both have essential roles during infection of the respiratory tract and sepsis. *Infect Immun*, **74**, 4014–20.

Martens, P., Worm, S. W., Lundgren, B., Konradsen, H. B., and Benfield, T. (2004). Serotype-specific mortality from invasive *Streptococcus pneumoniae* disease revisited. *BMC Infect Dis*, **4**, 21.

McCoy, S., and Pettigrew, M. (2003). Molecular epidemiology of *Streptococcus pneumoniae* mediated otitis media. *Front Biosci*, **8**, e87–93.

Meats, E., Brueggemann, A. B., Enright, M. C., *et al.* (2003). Stability of serotypes during nasopharyngeal carriage of *Streptococcus pneumoniae*. *J Clin Microbiol*, **41**, 386–92.

Mora, M., Bensi, G., Capo, S., *et al.* (2005). Group A *Streptococcus* produce pilus-like structures containing protective antigens and Lancefield T antigens. *Proc Natl Acad Sci USA*, **102**, 15641–6.

Nallapareddy, S. R., Singh, K. V., Sillanpaa, J., *et al.* (2006). Endocarditis and biofilm-associated pili of *Enterococcus faecalis*. *J Clin Invest*, **116**, 2799–807.

Obert, C., Sublett, J., Kaushal, D., *et al.* (2006). Identification of a candidate *Streptococcus pneumoniae* core genome and regions of diversity correlated with invasive pneumococcal disease. *Infect Immun*, **74**, 4766–77.

Oggioni, M. R., and Claverys, J. P. (1999). Repeated extragenic sequences in prokaryotic genomes: a proposal for the origin and dynamics of the RUP element in *Streptococcus pneumoniae*. *Microbiology*, **145**, 2647–53.

Oggioni, M. R., Memmi, G., Maggi, T., *et al.* (2003). Pneumococcal zinc metalloproteinase ZmpC cleaves human matrix metalloproteinase 9 and is a virulence factor in experimental pneumonia. *Mol Microbiol*, **49**, 795–805.

Orihuela, C. J., Gao, G., Mcgee, M., *et al.* (2003). Organ-specific models of *Streptococcus pneumoniae* disease. *Scand J Infect Dis*, **35**, 647–52.

Orihuela, C. J., Gao, G., Francis, K. P., *et al.* (2004a). Tissue-specific contributions of pneumococcal virulence factors to pathogenesis. *J Infect Dis*, **190**, 1661–9.

Orihuela, C. J., Radin, J. N., Sublett, J. E., Yu, J., and Tuomanen, E. I. (2004b). Microarray analysis of pneumococcal gene expression during invasive disease. *Infect Immun*, **72**, 5582–96.

Paterson, N. G., Riboldi-Tunicliffe, A., Mitchell, T. J., and Isaacs, N. W. (2006). Overexpression, purification and crystallization of a choline-binding protein CbpI from *Streptococcus pneumoniae*. *Acta Crystallogr Sect F Struct Biol Cryst Commun*, **62**, 672–5.

Pebody, R. G., Leino, T., Nohynek, H., *et al.* (2005). Pneumococcal vaccination policy in Europe. *Eur Surveill*, **10**, 174–8.

Pettigrew, M. M., and Fennie, K. P. (2005). Genomic subtraction followed by dot blot screening of *Streptococcus pneumoniae* clinical and carriage isolates identifies genetic differences associated with strains that cause otitis media. *Infect Immun*, **73**, 2805–11.

Pettigrew, M. M., Fennie, K. P., York, M. P., Daniels, J., and Ghaffar, F. (2006). Variation in the presence of neuraminidase genes among *Streptococcus pneumoniae* isolates with identical sequence types. *Infect Immun*, **74**, 3360–5.

Rosini, R., Rinaudo, C. D., Soriani, M., *et al.* (2006). Identification of novel genomic islands coding for antigenic pilus-like structures in *Streptococcus agalactiae*. *Mol Microbiol*, **61**, 126–41.

Sandgren, A., Sjostrom, K., Olsson-Liljequist, B., *et al.* (2004). Effect of clonal and serotype-specific properties on the invasive capacity of *Streptococcus pneumoniae*. *J Infect Dis*, **189**, 785–96.

Sandgren, A., Albiger, B., Orihuela, C. J., *et al.* (2005). Virulence in mice of pneumococcal clonal types with known invasive disease potential in humans. *J Infect Dis*, **192**, 791–800.

Schmidt, H., and Hensel, M. (2004). Pathogenicity islands in bacterial pathogenesis. *Clin Microbiol Rev*, **17**, 14–56.

Scott, J. R., and Zahner, D. (2006). Pili with strong attachments: Gram-positive bacteria do it differently. *Mol Microbiol*, **62**, 320–30.

Shakhnovich, E. A., King, S. J., and Weiser, J. N. (2002). Neuraminidase expressed by *Streptococcus pneumoniae* desialylates the lipopolysaccharide of *Neisseria meningitidis* and *Haemophilus influenzae*: a paradigm for interbacterial competition among pathogens of the human respiratory tract. *Infect Immun*, **70**, 7161–4.

Silva, N. A., McCluskey, J., Jefferies, J. M., *et al.* (2006). Genomic diversity between strains of the same serotype and multilocus sequence type among pneumococcal clinical isolates. *Infect Immun*, **74**, 3513–8.

Sjostrom, K., Spindler, C., Ortqvist, A., *et al.* (2006). Clonal and capsular types decide whether pneumococci will act as a primary or opportunistic pathogen. *Clin Infect Dis*, **42**, 451–9.

Sjostrom K., Blomberg C., Fernebro J., *et al.* (2007). Clonal success of piliated penicillin nonsusceptible pneumococci. *Proc Natl Acad Sci USA*, **104**, 12907–12.

Tai, S. S. (2006). *Streptococcus pneumoniae* protein vaccine candidates: properties, activities and animal studies. *Crit Rev Microbiol*, **32**, 139–53.

Takamatsu, D., Bensing, B. A., and Sullam, P. M. (2004). Four proteins encoded in the *gspB-secY2A2* operon of *Streptococcus gordonii* mediate the intracellular glycosylation of the platelet-binding protein GspB. *J Bacteriol*, **186**, 7100–11.

Takamatsu, D., Bensing, B. A., Cheng, H., *et al.* (2005a). Binding of the *Streptococcus gordonii* surface glycoproteins GspB and Hsa to specific carbohydrate structures on platelet membrane glycoprotein Ibα. *Mol Microbiol*, **58**, 380–92.

Takamatsu, D., Bensing, B. A., and Sullam, P. M. (2005b). Two additional components of the accessory *sec* system mediating export of the *Streptococcus gordonii* platelet-binding protein GspB. *J Bacteriol*, **187**, 3878–83.

Takamatsu, D., Bensing, B. A., Prakobphol, A., Fisher, S. J., and Sullam, P. M. (2006). Binding of the streptococcal surface glycoproteins GspB and Hsa to human salivary proteins. *Infect Immun*, **74**, 1933–40.

Telford, J. L., Barocchi, M. A., Margarit, I., Rappuoli, R., and Grandi, G. (2006). Pili in gram-positive pathogens. *Nat Rev Microbiol*, **4**, 509–19.

Tettelin, H., Nelson, K. E., Paulsen, I. T., *et al.* (2001). Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science*, **293**, 498–506.

Wani, J. H., Gilbert, J. V., Plaut, A. G., and Weiser, J. N. (1996). Identification, cloning, and sequencing of the immunoglobulin A1 protease gene of *Streptococcus pneumoniae*. *Infect Immun*, **64**, 3967–74.

CHAPTER 10

The Mobile Genetic Elements of *Staphylococcus aureus*

Richard P. Novick

10.1. INTRODUCTION

Like most eubacteria, *S. aureus* possesses a variety of mobile genetic elements (MGEs) that contribute in major ways to pathogenesis and its evolution. In addition to the typical MGEs carried by most bacteria, that is, prophages, transposons, and plasmids, *S. aureus* possesses two types of novel elements that have not been described for other bacteria, namely the superantigen-encoding pathogenicity islands and the resistance-encoding SCC*mec* elements. In this chapter, the general properties of these various MGEs are summarized, with special emphasis on the two novel types and on their contributions to pathogenesis and its evolution.

10.2. MOLECULAR GENETICS OF THE STAPHYLOCOCCAL MGEs

10.2.1. Plasmids

For a comprehensive review of plasmid origins and interactions, see Firth and Skurray (2006). Staphylococcal plasmids range in size from 1.2 to more than 100 kb; all known staphylococcal plasmids are circular duplex DNA, using either of the two standard modes of replication, theta and rolling circle (RC), with the latter being used principally by those of less than 10 kb, and the former by those larger, though this is only an approximate dividing line. As with all other plasmids, replication of staphylococcal plasmids is negatively autoregulated. For the known small RC plasmids, this is accomplished by *cis*-encoded antisense RNAs, sometimes with the assistance of small proteins. Theta plasmids of the pSK41/pGO1 family also appear to use an antisense mechanism.

10.2.1.1. Replication and its Regulation

All known staphylococcal plasmids encode replicon-specific initiator proteins (IPs) and, in all known cases, regulate their copy numbers by negatively controlling the rate of synthesis of these proteins. The most fully characterized plasmid replication systems in staphylococci are those of the small multicopy RC plasmids, of which pT181 and several very close relatives are prototypes (Khan, 2005). For these plasmids, each newly synthesized IP dimer results in a single replication event, according to the following scheme: As for other RC replicons, replication is initiated by the introduction of a nick in the leading strand at a specific site, the leading strand origin, which must be melted prior to nicking. The nick site is at the tip of a potential hairpin and is melted by the dimeric IP, using the free energy of supercoiling to drive unwinding. The IP remains attached to the 5′ end of the nicked leading strand by a phosphotyrosine bond. Extension of the 3′ end, catalyzed by the host replicase, displaces the 5′ end with the attached IP, which remains physically associated with the replication complex. At the end of a replication cycle, leading-strand extension continues for 10 nt beyond the nick site, sufficient to allow formation of the origin hairpin. This stops the extension and triggers a strand exchange reaction in which both the displaced and nascent leading strands are circularized and the IP is released with the extra 10 nt of the leading strand remaining attached to the active site tyrosine of one subunit. This oligonucleotide tail serves to inactivate the IP, thus preventing its re-utilization and ensuring the accuracy of replication regulation (Rasooly and Novick, 1993). Replication of the displaced monomeric leading strand involves a large palindromic sequence known as *palA*, which serves as the lagging- or single-strand replication origin (*sso*). The *sso* varies widely but is not replicon-specific, as its key feature is a sequence that, in its double-stranded configuration, can be recognized as a promoter and is used for the synthesis of a short primer by the host RNA polymerase (Rpo). This primer is extended by DNA polymerase to complete synthesis of the complementary strand, which is then re-circularized (Kramer *et al.*, 1998; Birch and Khan, 1992). *sso* function is orientation- but not location-specific: If it is in the wrong orientation, plasmid replication is sharply curtailed and single-stranded monomers representing the displaced leading strand accumulate (Gruss *et al.*, 1987).

IP synthesis for plasmids of this type is tightly controlled by an unstable antisense RNA that is transcribed from the untranslated leader region of the IP gene. This countertranscript engages in loop-loop pairing with the nascent sense transcript, causing the formation of a hairpin immediately 5′ to the

IP translation start site. This aborts IP translation by causing attenuation of the IP mRNA, the probability of which is determined by the intracellular countertranscript concentration (Novick *et al.*, 1989). As individual plasmid molecules are chosen for replication from a multicopy pool at random times during the cell cycle, disparities in copy number inevitably develop within individual cells. This regulation system rapidly and automatically corrects these disparities: Any excess in the number of copies causes a corresponding increase in countertranscript concentration, which reduces the rate of IP synthesis. Conversely, a decrease in the number of copies is rapidly followed by a decrease in concentration of the short-lived countertranscript, resulting in an increase in IP synthesis (Highlander and Novick, 1990).

Identical plasmid replicons are always incompatible – that is, they cannot stably coexist in a single strain. Incompatibility among staphylococcal plasmids, so far as is known, is strictly dependent on copy numbers. Although it is possible with multicopy plasmids to introduce two differentially marked but otherwise identical plasmids into a single strain, balanced heterozygosity cannot be maintained since individual molecules are chosen for replication and segregation at random from the common copy pool. This results in the development of disparities within individual cells, which cannot be corrected since both plasmids are subject to the same regulation, eventually resulting in segregation of organisms containing only one of the two. The rate of this segregation can be predicted by simple stochastic methods (Novick, 1987). Sharing of either replication-origin or copy-control specificities causes incompatibility, but segregation rates are slower than when both specificities are shared (Highlander and Novick, 1990).

Mobility is based primarily on conjugation, a property of the widespread pSK41 family, which can mobilize small plasmids that possess transfer origins and ancillary proteins forming a relaxation complex (Projan and Archer, 1989). Mating requires intimate cell-cell contact and is seen primarily on solid substrata, though it has been reported to occur in fluids such as blood, serum, and urine. The conjugation (*trs*) determinant of the pGO1/pSK41 plasmids (Thomas and Archer, 1989) is an approximately 14-kb operon that, interestingly, is flanked by direct repeats of the common insertion sequence, IS*257*/*431*. The *trs* operon, however, does not include either *oriT,* the site of transfer initiation, or the gene encoding the site-specific nicking enzyme (Nes) that is responsible for the initiation of transfer. The entire region may originally have been contiguous, interrupted by the insertion of an IS*257*-flanked segment (Berg *et al.*, 1998). pSK41 has been sequenced (Berg *et al.*, 1998), and a linear map is presented in Figure 10.1. The related streptococcal conjugative plasmid pIP501 has a very broad host range

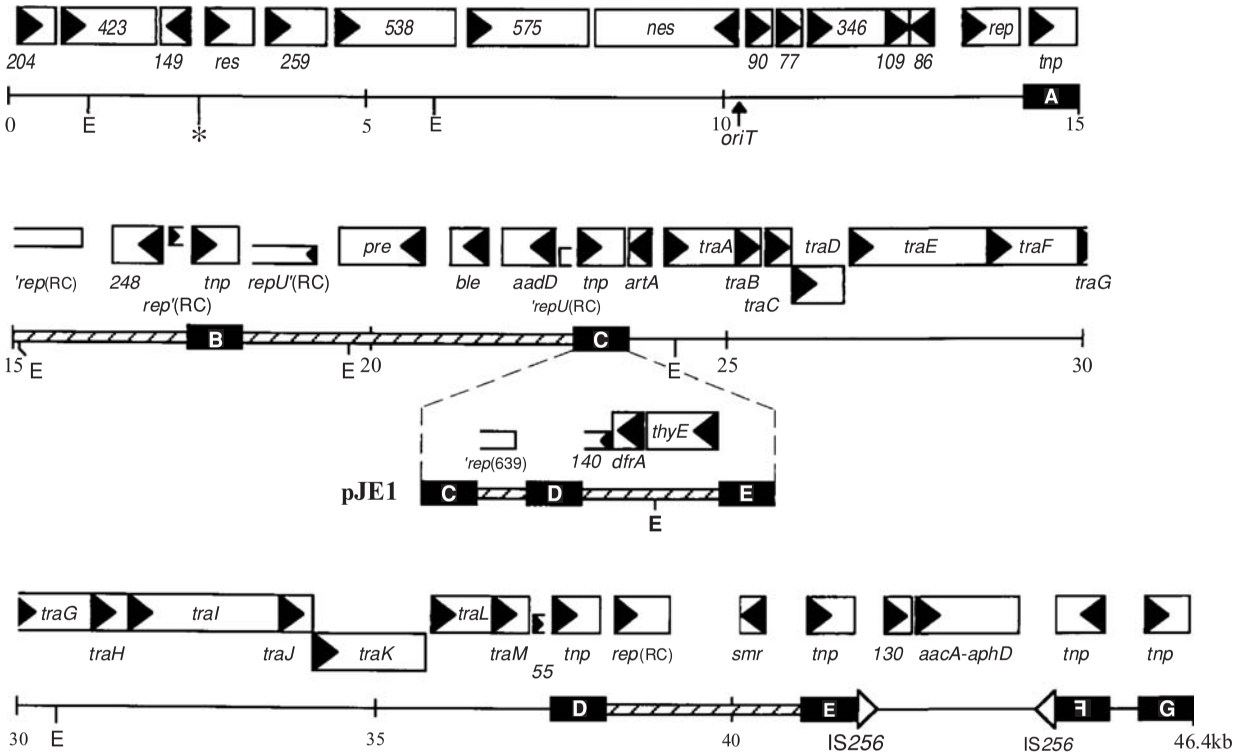

423
538
575
nes
346
rep
204
149
res
259
90
77
109
86
tnp
A
0
E
5
E
10
oriT
15
'rep(RC)
248
rep'(RC)
tnp
repU'(RC)
pre
ble
aadD
tnp
'repU(RC)
artA
traA
traB
traC
traD
traE
traF
traG
B
C
15
E
E
20
E
25
30
thyE
'rep(639)
140
dfrA
pJE1
C
D
E
E
traG
traI
traH
traJ
traK
traL
traM
55
tnp
rep(RC)
smr
tnp
130
aacA-aphD
tnp
tnp
D
E
F
G
30
E
35
40
IS256
IS256
46.4kb

Figure 10.1. Genetic organization of pSK41. Genes are represented by open boxes with names shown within or below and with arrowheads indicating the direction of transcription; smaller open boxes represent interrupted genes. Copies of IS*257* are denoted by black boxes containing the element's designation (A through G), whereas inverted copies of truncated IS*256* elements are indicated by open triangles. Hatched segments indicate integrated small plasmids. RC plasmid replication-initiation genes are suffixed (RC) to differentiate them from the probable theta-mode rep gene of pSK41. Kilobase coordinates are shown at bottom, as are the positions of *Eco*RI restriction sites (E) and *oriT* (vertical arrow). The position in pSK41 corresponding to the insertion site of the Tn*552*-like transposon in pUW3626 is indicated by an asterisk (*). The position and genetic organization of the additional DNA segment present in pJE1 are also indicated; the *tnp* genes of the IS*257* elements have been omitted for clarity. The pSK639-like *rep* gene remnant is suffixed (639) to differentiate it from pSK41 *rep*. Genetic nomenclature is as follows: numbers, size in codons of deduced ORF of unknown function; *res*, resolvase; *oriT*, origin of conjugative transfer; *nes*, *oriT* nickase; *rep*/*repU*, replication initiation; *tnp*, transposase; pre, recombinase; *artA*/*traA*-M, transfer-associated genes; *thyE*, thymidilate synthetase. (Reproduced from Berg *et al.*, 1998, with kind permission of the publisher.)

among Gram-positive bacteria (Macrina and Archer, 1993), suggesting that conjugative plasmids have an important role in horizontal gene transfer in Gram-positive as well as in Gram-negative bacteria.

Several of the small RC plasmids exist as relaxation complexes (*rlx*), analogous to those seen with plasmids of the ColE1 family (Novick, 1976), consisting of three proteins (MobA, B, and C) and a specific nicking site, corresponding to *oriT*. These can be mobilized at high frequency by pGO1 and other conjugative plasmids and are frequently transferred without the mobilizing plasmid (Projan and Archer, 1989). Deletion of an 8.6-kb segment of plasmid pMB494 has activated a previously quiescent conjugation function. The resulting plasmid, pXU12, can transfer not only pXU12 itself and the *rlx* plasmids, but also several small plasmids lacking any relaxation complex, all at very high frequency (Udo and Jacob, 1998). A site-specific recombination system known as the *pre*-RS$_A$ system (Gennaro *et al.*, 1987), which is responsible for the formation of plasmid cointegrates, may be responsible for the buildup of multiresistance plasmids and for the conjugative transfer by pXU12 of small plasmids that lack a transfer origin; such transfer would theoretically be by conduction, though this has not actually been demonstrated to date. Chromosomal genes could be transferred by a conjugative plasmid when both chromosome and plasmid carried a copy of Tn*551* (Stout and Iandolo, 1990). The mechanism of this is unclear, since there was neither regional nor orientation specificity of the genes transferred.

Plasmids are the major repository of staphylococcal resistance genes and are largely responsible for their horizontal spread (Lyon and Skurray, 1987). Small RC plasmids usually carry one, sometimes two resistance determinants; larger plasmids have been identified that carry six or seven, including resistance to heavy metals and disinfectants as well as to antibiotics. Resistance genes carried by staphylococcal plasmids and other MGEs are listed in Table 10.1. On the other hand, staphylococcal plasmids and transposons carry virulence genes extremely rarely, with enterotoxin D and exfoliatin B being the only known examples – unless one considers bacteriocins, which are carried only by plasmids, as virulence factors. Remarkably, many *S. aureus* strains secrete short hydrophobic peptides similar or identical to the mating pheromones of *Streptococcus faecalis,* enabling them to serve as recipients for *S. faecalis* plasmids (Nakayama *et al.*, 1996). Although there is no evidence that these are involved in intraspecific mating in staphylococci, they are presumably responsible for several recent transfers from *S. faecalis* to *S. aureus* of conjugative plasmids carrying Tn*1546*, a conjugative transposon containing the notorious *vanHAX* (vancomycin resistance) operon (Weigel *et al.*, 2003). This transfer has generated the dangerous VRSA strains that

Table 10.1. *Resistances carried by staphylococcal MGEs*

Resistance	Gene	MGE	Size (kb)
Streptomycin	*aadE*	pS194	4.5
Neomycin/Kanamycin	*aadD*	pUB110	4.5
Bleomycin	*ble*	PUB110	4.5
Spectinomycin	*spc*	Tn*554*	5.4
Erythromycin (MLS)	*ermA*	Tn*554*	5.4
Erythromycin (MLS)	*ermB*	Tn*551* (pI258)	5.3 (28)
Erythromycin (MLS)	*ermC*	pE194, pE5	2.2–3.2
Chloramphenicol	*cat*	pC194, pC221, etc	3.8–4.5
Gentamicin	*aacA-aphA*	Tn*4001* (pSK41)	4.7 (45)
Tetracycline	*tetA(K)*	pT181	4.4
Tetracycline	*tetM*	Tn*5801*	11
Vancomycin	*vanHAX*	Tn*1546* (pLW1043)	10.8
Methicillin	*mecA*	SCC*mec*	20–60
Mupirocin	*mupA*		
β-lactam	*blaZ*	Tn*552* (pI524, etc)	6.1 (31)
Trimethoprim	*dfrA*	Tn*4003*	4.7
Arsenate	*asa*	pI524, etc	31
Arsenite	*asi*	pI524, etc	31
Cadmium	*cadA*	pI524, etc	31
Cadmium	*cadB*	pII147	33
Cadmium	*cadC*		
Cadmium	*cadD*		
Mercury	*merA,B*	Tn*4004* (pI258, etc)	7.8 (28)
Quaternary amines	*qac*	pWG302	3.2

are threatening to eliminate vancomycin as the antibiotic of last resort for life-threatening staphylococcal infections.

10.2.2. Transposons

Staphylococci have their fair share of transposons and transposon-like elements (for review, see Firth *et al.*, 2000), the former represented principally by Tn*551*, Tn*552*, and Tn*4524*. Initially found on plasmids, these are similar to the *E. coli* prototype Tn*3* and carry single resistances to MLS, β-lactam, and aminoglycoside antibiotics, respectively. Transposons of this class consist

essentially of a transposase gene, a resolvase gene, a resistance gene, and short terminal inverted repeats. Tn*552* has, in addition, an invertible segment whose inversion is catalyzed by the transposon-coded resolvase. In addition to normal transposition, Tn*552* moves reversibly and at high frequency by resolvase-catalyzed recombination between a copy of one of the repeats located on the chromosome of strain 9789, the strain in which it was originally found, and one of the inverted repeats located on the resident plasmid, pI9789. These transposons typically transpose to ordinary target sites at frequencies in the 10^{-6} to 10^{-7} range, have strong regional or site preferences, and generate 5 or 6 nt direct repeats. Tn*551* is virtually identical to Tn*917* from *S. faecalis* (Wu *et al.*, 1999) and must certainly have been transferred from one to the other or to both from a common progenitor, thus representing the first proof of interspecific transfer between streptococci and staphylococci of a specific MGE. *S. aureus* also possesses a high-frequency site-specific transposon, Tn*554* (Krolewski *et al.*, 1981), which carries MLS and aminoglycoside resistances and is similar to Tn*7* of *E. coli*. Tn*554* transposes at a frequency approaching 100% to its primary site and at much lower frequencies to secondary sites located elsewhere on the chromosome, on plasmids, and on other mobile genetic elements, especially SCC*mec* elements. A wide variety of transposon-like elements have been identified by sequencing of genomes, including plasmids and chromosomal islands. Although some of these appear to be intact, functionality has yet to be demonstrated directly for any of them. Included in this group are several insertion sequence (IS)-like elements that have clearly been mobile as they occupy a variety of sites, including intragenic sites, and are evidently responsible for a variety of rearrangements (Firth and Skurray, 2006). Also included is at least one apparently intact conjugative transposon; though its functionality has not been demonstrated, staphylococci are well established as recipients of conjugative transposons from other organisms, including Tn*916* and Tn*1546* from *S. faecalis*, which are functional in *S. aureus*. In Table 10.2 are listed the known transposons and apparently intact transposon-like elements in staphylococci.

10.2.3. Prophages

The vast majority of *S. aureus* strains harbor prophages, commonly three or more. All those thus far known are linearly inserted into the chromosome and are flanked by a directly repeated attachment site sequence of 15–20 nt. Staphylococcal phages belong to the two well-known classes based on packaging mechanisms, namely the *cos* and *pac* types, of which the latter

Table 10.2. *Staphylococcal insertion sequences and transposons*[*]

Element	Size (kb)	TIR (bp)[†]	Target duplic (bp)	Location(s)[‡]
IS*256*	1.3	26	8	C, P
IS*257/431*	0.8	27	8	P, CI
IS*1181*	2.0	23	8	C
IS1182	1.9	33	8	C
IS*1272*	1.9	16	Unknown	C
IS*1293*	1.3	26	Unknown	P
Tn*551*	5.3	40	5	P
Tn*552*[§]	6.1	116	6/7	C, P
Tn*554*[‖]	6.7	Absent	Absent	C, P, CI
Tn*1546*	10.8	38	5	P
Tn*3854*	4.5	Unknown	Unknown	P
Tn*4001*[¶]	4.7	IS*256* (I)[#]	8	C, P
Tn*4003*	4.7	IS*257* (D)[#]	8	P
Tn*4004*	7.8	IS*257* (D)	8	P
Tn*5404*	16	116	6	C
Tn*5405*	12	IS*1182* (I)	8	C
Tn*5406*	5.5	Absent	Absent	C, P
Tn*5801*	25.8	Absent	11	C

[*] Adapted from Firth and Skurray (2006). See Paulsen *et al.* (1997) and Firth and Skurray (1998) for references.

[†] TIR, terminal inverted repeat.

[‡] C, chromosome; P, plasmid; CI, chromosomal island.

[§] Tn*3852*, Tn*4002* and Tn*4201* are likely to be similar or identical to Tn*552*.

[‖] Tn*3853* is likely to be similar or identical to Tn*554*.

[¶] Tn*3851* and Tn*4031* are likely to be similar or identical to Tn*4001*.

[#] (I), inverted orientation; (D), direct orientation

are probably all generalized transducing phages. Specialized transduction, seen with *cos* phages such as coliphage λ, has yet to be demonstrated in staphylococci, although it has been suggested to be responsible for the development of converting phages (see later discussion). All known staphylococcal prophages, like those in other bacteria, are SOS-inducible, always by mitomycin C, but not always by ultraviolet radiation. Staphylococcal phages have typical tailed virions with polyhedral or cylindrical heads, but, interestingly, the *S. aureus* phages do not seem to utilize specific receptors, being able to

attach to a non-specific component of the ribitol teichoic acid cell wall polymer (Chatterjee, 1969). Therefore, host specificity in *S. aureus* is generally based on post-adsorption phenomena, such as the presence or absence of specifically required host factors, classical prophage immunity, restriction-modification functions, and phage-specific interference mechanisms other than classical immunity. A phage typing system, based on host specificity, was developed in the 1950s and was widely used for many years; however, because the basis of host specificity was never analyzed and because phage type types tend to rather labile, the system has been largely abandoned in favor of various genotyping systems.

Primary genome sequences are now available for at least 50 different temperate staphylococcal phages, Some 30 determined by direct sequencing of purified phage DNAs (Kwan *et al.*, 2005) and 20–30 others included in the nine sequenced genomes. Although staphylococcal phage genomes have the modular organization typical of phages in general, they are highly mosaic, being composed of assorted segments assembled almost at random – indicating highly promiscuous recombination that does not always seem to be based on clearly identifiable sequence homology. This suggests the existence of recombination functions that can operate on very short homologous sequences, or possibly on non-homologous sites.

Staphylococcal phages have a major role in pathogenesis, being the carriers of a large variety of virulence factors, including toxins, adhesins, anti-phagocytic factors, and pathogenic enzymes, causing *lysogenic conversion*. Unlike most prophage genes, which are not expressed in the prophage state, toxin and other virulence genes are well expressed; indeed, some are induced by SOS induction of the prophage, causing overexpression of the toxin gene and enhancing pathogenesis, as well as spread of the toxin genes (Waldor and Friedman, 2005). This is of particular concern in connection with fluoroquinolones and other gyrase inhibitors; since these drugs are SOS inducers, treatment with fluoroquinolones of a disease in which these genes are involved may have the serious adverse consequences associated with increased toxin production (Zhang *et al.*, 2000). Virulence genes known to be carried by staphylococcal prophages are listed in Table 10.3. Among these, the toxins are always located at one end of the prophage, which has suggested to some that they may have been incorporated into the phage genome by an aberrant excision mechanism such as that by which the defective special λ transducing phages, λ*dg* and λ*db*, were generated. However, this seems rather unlikely because most of the staphylococcal converting phages are not defective and the known phage *att* sites are not adjacent to any of the toxin genes – indeed, the toxin genes carried by these phages do not occur in the

Table 10.3. *Staphylococcal toxins and other virulence factors carried by MGEs*

Condition or function	Toxin/virulence factor	MGE	Size (kb)
Toxic shock syndrome (TSS)	TSST-1	SaPI	~15 kb
Food poisoning, TSS	Enterotoxin A, P	Prophage	~45 kb
Food poisoning, TSS	Enterotoxin B	SaPI	~15 kb
Food poisoning, TSS	Enterotoxin C	SaPI	~15 kb
Food poisoning, TSS	Enterotoxin D	Plasmid	~30 kb
Food poisoning, TSS (?)	Enterotoxin J	Plasmid	~30 kb
Food poisoning, TSS (?)	Enterotoxin K	SaPI	~15 kb
Food poisoning, TSS (?)	Enterotoxin L	SaPI	~15 kb
Food poisoning, TSS (?)	Enterotoxin M	SaPI	~15 kb
Food poisoning, TSS (?)	Enterotoxin Q	SaPI	~15 kb
Scalded skin syndrome	Exfoliatin A	Prophage	~45 kb
Scalded skin syndrome	Exfoliatin B	Plasmid	~30 kb
Necrotizing pneumonia	Leukocidin (PVL)	Prophage	~45 kb
Fibrinolysis	Staphylokinase	Prophage	~45 kb
Anti-complement	CHIPS	Prophage	~45 kb
Adhesin	BAP	SaPI	28 kb

staphylococcal genomes except as phage components, suggesting that both toxin genes and their carrier phages may have been imported from other species – but if so, by an entirely unknown mechanism, given the very limited adsorption specificity of these phages, plus the rarity of these genes in other staphylococci. Thus, the origins of both the toxin genes and of their carrier phages remain a great mystery. In one case, that of the converting phages carrying *sak,* CHIPS, and *sea,* these genes are grouped, appear to have been incorporated into the phage genome en bloc, and may represent a small pathogenicity island. In this case, the apparently inserted DNA could be as much as 9 kb in length – a size that would be expected to generate a defective prophage. Nevertheless, these phages are generally not defective, though they grow extremely slowly. An interesting possibility would be that their capsids are sufficiently flexible in size to accommodate such an increment in the size of the phage genome. Flexibility in capsid size has, in fact, been observed in connection with the SaPIs (see later discussion).

A special feature of certain staphylococcal converting phages is that their *att* sites are located within genes belonging to the virulon, namely *hlb*

(encoding β-hemolysin; Coleman *et al.*, 1989) and *geh* (encoding the major staphylococcal lipase; Lee and Iandolo, 1986). Insertion into these sites, known as 'negative conversion', was the first case of lysogenic conversion discovered (Winkler *et al.*, 1965). Though prophage *att* sites within coding sequences are seen in other organisms, these are generally in genes of unknown function and have not received any attention; thus the negative conversion phenomenon in *S. aureus* is often (and incorrectly) regarded as unique. It is also notable that the phages that insert into *hlb* are the ones just mentioned that carry some combination of *sak, sea,* and CHIPS, so that their presence causes positive as well as negative conversion (Coleman *et al.*, 1989).

A particularly troublesome case of lysogenic conversion in *S. aureus* is represented by phages that carry the genes for Panton-Valentine leukocidin (PVL) (Zou *et al.*, 2000; Narita *et al.*, 2001). PVL is a well-characterized staphylococcal toxin and has long been known to be involved in relatively serious but not life-threatening staphylococcal skin infections, such as furunculosis (Lina *et al.*, 1999). Recently, PVL has been identified as the causative toxin in a newly recognized fulminant necrotizing pneumonitis that carries a mortality rate well in excess of 50% in most series (Gillet *et al.*, 2002). Since a variety of clonal lineages have been found to produce this toxin (Nimmo *et al.*, 2006; Johnson *et al.*, 2006), it is clear that the carrier phages are spreading rather widely; moreover, the newly predominant methicillin-resistant *Staphylococcus aureus* (MRSA) lineages, such as USA300, that are responsible for serious outbreaks of staphylococcal infections, including necrotizing pneumonitis (King *et al.*, 2006), carry a PVL-encoding prophage and produce this toxin.

10.2.4. Pathogenicity Islands

Pathogenicity islands have recently been recognized as the repository of virulence genes in many organisms (Groisman and Ochman, 1996; Hochhut *et al.*, 2005). Although there is little doubt that they have been acquired by horizontal transmission and that their acquisition is clearly responsible for the conversion of weakly pathogenic or non-pathogenic enteric organisms into major pathogens, their mobility has been, until very recently, difficult or impossible to demonstrate. This has suggested that following acquisition, they have become transfer defective owing to deletions, mutations, or other genotypic rearrangements.

The situation in staphylococci is somewhat different. The first staphylococcal pathogenicity islands (SaPIs) identified (SaPIs 1, 2, and 3) were discovered because of their carriage of the genes for toxic shock syndrome

toxin-1 (TSST-1) and other superantigens. They were soon found to be highly mobile, owing to their relationship with certain phages (Lindsay *et al.*, 1998). The SaPIs are inserted at specific chromosomal sites (att_S) and each is always in the same orientation. With the exception of SaPIbov2 (27 kb), the functional ones are approximately 14–17 kb; a highly degenerate one, of 3.14 kb, is present in five of the sequenced *S. aureus* genomes. The SaPIs, of which complete sequences are known for 15, including two in non-*aureus* staphylococci (see later discussion), form a highly coherent family with conserved functional and genetic organization. Several other putative pathogenicity islands have been described on the basis of genome sequencing and other studies; these are 25–30 kb in length and encode sets of enterotoxin-like, protease-like, or lipase-like proteins. restriction-modification systems, and/or transposon remnants (for example, see Kuroda *et al.*, 2001). They are rather widespread among *S. aureus* strains, mobility has not been demonstrated, and their designation as horizontally transferred chromosomal islands is inferential. They are not detailed here, since this chapter is concerned exclusively with mobile elements. Although coliphage phage P4 (reviewed in Lindqvist *et al.*, 1993; Bensing *et al.*, 2004) and *Sulfolobus* plasmid pSSVx (Arnold *et al.*, 1999) share several properties with the SaPIs such as functional dependence on specific phages, such elements have yet to be described in other Gram-positive bacteria.

10.2.4.1. Nomenclature

'SaPI' (staphylococcal pathogenicity island) was initially used to designate chromosomally located mobile elements in staphylococci on the basis of their superantigen content (Lindsay *et al.*, 1998). This designation was reinforced when it became clear that the basic core genome organization of these elements was highly conserved and unique (see later discussion). For this reason, it is proposed to continue the use of 'SaPI' for *S. aureus* genomic islands that belong to this clearly defined family, rather than the nomenclature proposed by Baba *et al.* (2002), which lumps all known or putative genomic islands, ignoring the uniqueness of the SaPIs while attempting to portray the SCC*mecs* as unique by excluding them from the general category of genomic islands. For the purpose of continuity, the Baba *et al.* designations are indicated parenthetically. Since insertion site specificity is arguably the defining feature of mobile elements such as these, it is suggested that the designation 'SaPI' be taken to imply site specificity. Previously described SaPIs, including those identified by Kuroda *et al.* (2005) as well as by ourselves and by others (Diep *et al.*, 2006), would retain their original designations. We suggest that newer elements belonging to the SaPI category be designated SaPIn,

where 'n' refers to the numerical order of discovery, or to a strain number. We therefore propose to redesignate SaGlm (**Type II** νSa3) as SaPIm4, and suggest SaPI122 for the newly described second SaPI in strain RF122 (GenBank NC_007622) and SaPI1028 for the element in NY940 initially described as a prophage (Kwan *et al.*, 2005). We also suggest SaPImw2 for the SaPI in strain mw2, also listed as '**Type II** νSa3' (Baba *et al.*, 2002) and SaPI6Δ for the 3.14 kb SaPI remnant present at the 44′ site in five of the sequenced genomes and designated as '**Type II** νSa4'. Since this remnant is present in five of the nine sequenced genomes, a unique strain-specific designation seemed inappropriate so the next available number, '6,' has been used. The recently published sequences of *S. haemolyticus* (Takeuchi *et al.*, 2005) and *S. saprophyticus* (Kuroda *et al.*, 2005) each contain a SaPI, whereas neither of the two known *S. epidermidis* genomes (Zhang *et al.*, 2003; Gill *et al.*, 2005) contains one. The element in *S. haemolyticus,* νSh2, would be re-designated ShPI2, and νSs15305 from *S. saprophyticus* strain 15305, SsPI15305.

10.2.4.2. Superantigen and Other Accessory Genes

Staphylococcal superantigens have been found, thus far, only in association with mobile genetic elements, most frequently SaPIs. The biological basis for this is entirely mysterious but is consistent with the observation that superantigens in other organisms, including not only group A streptococci but also mice, are carried by mobile genetic elements including phages and retroviruses such as murine tumor virus (Marrack *et al.*, 1993). The toxin most frequently encoded by SaPIs is toxic shock syndrome toxin-1 (TSST-1), which has been found, thus far, only associated with SaPIs that are located at three of the six known SaPI *att* sites; these are therefore exclusively responsible for menstrual TSS and are most commonly carried by *agr* group III strains (Musser *et al.*, 1990). Superantigen and other virulence genes are located at either end of a SaPI, and some are flanked by non-matching sequences, suggesting that they have been inserted into pre-existing genetic units by non-homologous recombination events. A particularly striking case is that of *tst* (encoding TSST-1) and *seb* (encoding enterotoxin B), which are inserted in opposite orientations at precisely the same site in SaPI1 and SaPI3, respectively. The sequences flanking *tst* and *seb* are also present in SaPI5, but do not contain any insertion.

Many of the SaPIs encode two or three superantigen toxins. Diverse additional accessory genes (i.e., unrelated to the SaPI life cycle) with possible roles in virulence or antibiotic resistance are present in some. Thus several carry the *ear* gene, encoding a penicillin binding protein homolog that determines penicillin resistance in *E. coli* (but not in *S. aureus*); others

carry a homolog of the *E. coli* ferrichrome ABC transporter, *fhuD*; one carries an exfoliatin A (*eta*) homolog; one carries *bap*, encoding a ~2000-kDa adhesin; one carries a multidrug export homolog (*mdr*); and finally, one carries homologs of aminoglycoside adenyl transferase (*aad*) (streptomycin resistance) and of a metallothiol transferase (*fosB*) (fosfomycin resistance). In most cases, however, these genes have been identified only by sequence similarity; their functionality has not been established.

10.2.4.3. SaPIs and Their Genomes

All of the sequenced *S. aureus* strains, with the exception of MRSA252, carry at least one SaPI element, Since these strains are largely unrelated, belonging to three of the four *agr* groups, it may be inferred that the vast majority of *S. aureus* strains contain SaPIs and that a considerable proportion do not carry any superantigen or other discernable accessory gene. Details of sizes, attachment site sequences, and so forth are listed in Table 10.4, with designations as noted earlier (SaPIm1 and SaPIn1 are essentially identical, as are the five SaPI6Δs, and are not listed separately).

In general, the SaPIs form a coherent group with a highly conserved set of core genes; among these, the integrase the Rep protein (usually annotated as helicase-like or primase-like) and the terminase small-subunit homolog are always present and there is always a characteristic point of transcriptional divergence, near the integrase gene, flanked by two open reading frames (ORFs) encoding transcriptional regulatory proteins. This conserved core genome, transcriptional organization, and mobility mechanism (described in more detail later) serves to separate the SaPIs from all other mobile or potentially mobile elements among the staphylococci. Additionally, most of the other SaPI-specific ORFs do not have close orthologs anywhere in the staphylococcal genomes or elsewhere in the database.

A diagrammatic comparison of the sequenced SaPIs is presented in Figure 10.2, showing the conservation of overall organization and homologies based on similarities to SaPI1. Genes with over 50% sequence similarity to the corresponding SaPI gene at the amino acid level are shown as dark stippled symbols; those clearly homologous but with less than 50% similarity are shown in light stippled or open symbols. This organization is characterized by two major groups of genes that are divergently transcribed, with the ORFs flanking the divergence shown in dark gray. These core SaPI genes have been partially characterized in SaPI and SaPIbov1 (P. Barry, *et al.*, and C. Ubeda, *et al.*, unpublished). Within the rightward group is a set of highly conserved genes including *ter*, a close homolog of the terminase small subunit of several phages from Gram-positive bacteria that has predictably been found to

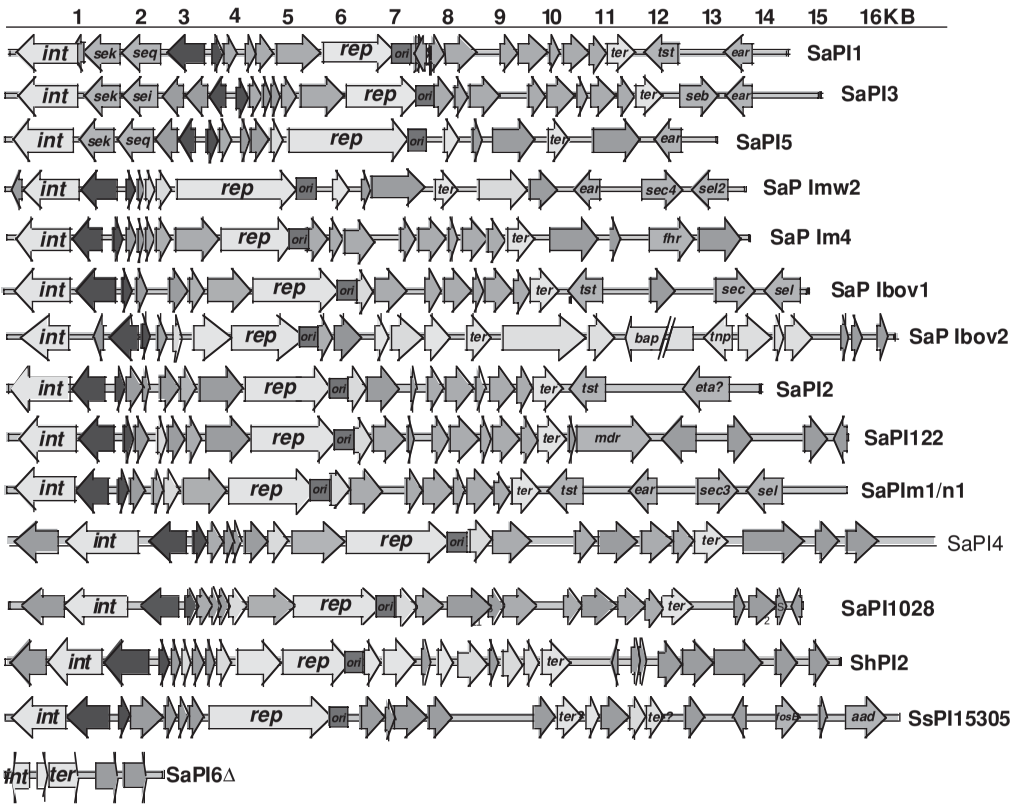

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16KB
int sek seq rep ori ter tst ear
SaPI1
int sek sei rep ori ter seb ear
SaPI3
int sek seq rep ori tet ear
SaPI5
int rep ori tet ear sec4 sel2
SaP Imw2
int rep ori ter fhr
SaP Im4
int rep ori ter tst sec sel
SaP Ibov1
int rep ori ter bap tnp
SaP Ibov2
int rep ori ter tst eta?
SaPI2
int rep ori ter mdr
SaPI122
int rep ori ter tst ear sec3 sel
SaPIm1/n1
int rep ori ter
SaPI4
int rep ori ter
SaPI1028
int rep ori ter
ShPI2
int rep ori ter? ter? osb aad
SsPI15305
int ter
SaPI6Δ

Figure 10.2. Comparison of SaPI genomes. Genomes are aligned according to the prophage convention with the integrase gene at the left end. Patterns for the ORFs are as follows: Genes that are functionally conserved in all SaPIs (*int*, *rep*, *ori* (rectangle) and *ter*) are shown black (sequence similarity varies widely); genes that mark the transcriptional divergence are shown in dark gray; most, or all of these, encode transcriptional regulators; genes encoding toxins and other potential virulence or resistance factors (stippled symbols) ; genes of unknown function involved in the SaPI life cycle are patterned according to their similarity to the corresponding genes in SaPI1: >70%, dark stippled symbols; 50–70%, mid stippled symbols; 20–50%, light stippled symbols; >20%, open symbols. The sequence illustrated here for SaPI 1028 is a permutation of the published sequence. *aad,* aminoglycoside transacetylase; *bap,* bovine adhesion protein; *ear, E. coli* ampicillin resistance; *eta,* exfoliatin A; *fosB,* fosfomycin resistance; *int*, integrase; *mdr,* multidrug export protein; *rep,* initiator protein (annotated as 'helicase' or 'primase'); *sec, sek, sel, seq,* enterotoxins C, K, L, Q, respectively; *ter,* terminase small subunit; *tnp,* transposase.

Table 10.4. *The SaPI family*

Element (strain)	Size, kb (comment)	Inducing phages	*att* site core (location)	Virulence genes	Orient.	References
SaPI4 (MRSA252)	15.1	Endogenous prophage	AAAGAAGAACAATAATAT (~ 8′)	None	+	Holden *et al.*, 2004; A. Subedi and RPN, unpublished data
SaPI1028 (NY940)	15.6	Endogenous prophage	AAAGAAGAACAATAATAT (~ 8′)	None	+	Khan, 2005; A. Subedi and RPN, unpublished data
SaPIbov1	15.8	ϕ11 ϕ1477,	TAATTATTCCCACTCAAT (~ 9′)	*tst, sel, sek*	+	Fitzgerald *et al.*, 2001
(RF122 νSa2)	(excised spontaneously)	80α, ϕ69	Ubeda *et al.*, 2003	Ubeda, *et al.*, in preparation		
SaPIbov2 (V329)	27, (excised spontaneously)	80α, ϕ69	TAATTATTCCCACTCGAT (~ 9′)	*bap*	+	Baba *et al.*, 2002); Ubeda *et al.*, in preparation

SaPIm4 (mu50: SaGIm νSa3 **Type I**)	14.4	Endogenous prophage	TCCCGCCGTCTCCAT (~ 18′)	*fhuD*	+	(Baba *et al.*, 2002; A. Subedi and RPN, unpublished data
SaPImw2 (mw2: νSa3 **Type II**)	14.4	Endogenous prophage	TCCCGCCGTCTCCAT (~ 18′)	*ear, sel2, sec4*	+	Takeuchi *et al.*, 2005; A. Subedi and RPN, unpublished data
ShPI2 (*S. haemolyticus*, νSh2)	16.6	ND*	TCCCGCCGTCTCCAT (48′)†	None	−	Lindsay *et al.*, 1998
SaPI1 (RN4282 νSa1)	15.2	80α, ϕ13 ϕ69 *sek, seq*	TTATTTAGCAGGAATAA (~ 19′)	*ear, tst*	+	Novick *et al.*, 2001; Ubeda, *et al.*, in preparation
SaPI3 (COL νSa1)	15.6	29 (?) ϕ69	TTATTTAGCAGGAATAA (~ 19′)	*ear, seb, sel, sek*	+	Diep *et al.*, 2006; Ubeda *et al.*, in preparation
SaPI5 (USA300)	14.0	ND	TTATTTAGCAGGAATAA (~ 19′)	*ear, sek, seq*	+	Kuroda *et al.*, 2001

(*cont.*)

Table 10.4 *(continued)*

Element (strain)	Size, kb (comment)	Inducing phages	*att* site core (location)	Virulence genes	Orient.	References
SaPIn1(n315) SaPIm1(mu50) (νSa4 **Type I**)	(Identical) 15	ϕ69	GTTTTACCATCATTCCCGGCAT (~ 44′)	*tst, sel, sec3*	–	Kuroda *et al.*, 2001; Ubeda *et al.*, in preparation
SaPI2 (RN3984)	14.7	80, 80α	ATTTTACATCATTCCTGGCAT (~ 44′)	*tst, eta*	–	Ruzin *et al.*, 2001; A. Subedi and RPN, unpublished data
SaPI122 (RF122)	17.9	Endogenous	GTTTTACATCATTCCTGGCAT (~ 44′)	*mdr*	–	GenBank NC_007622
SaPI6Δ (8325, COL, USA300, MSSA476, mw2: νSa4 **Type II**)	(identical) 3.14 (excised spontaneously)	ND	GTTTTACCATCATTCCCGGCAT, GTTTTACATCATTCCTGGCAT (~ 44′)	None RPN, unpublished	–	Baba *et al.*, 2002
SsPI15305 (*S. saprophyticus* 15305, νSs15305)	17.0	ND	CGAGGGGACTAATAAGT, CGAGGGGATTAATAAGT (47′)	*aad, fosB*	–	Kuroda *et al.*, 2005

* ND, no data.

† ShPI2 is located 180° away from the other SaPIs having the same *att* core sequence, owing to the major chromosomal inversion that has been documented in the *S. haemolyticus* genome (Takeuchi *et al.*, 2005).

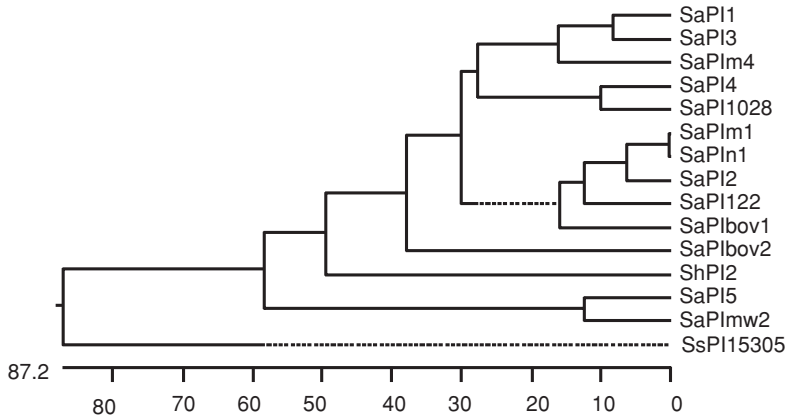


Figure 10.3. SaPI tree.

be required for encapsidation (C. Ubeda *et al.*, in preparation). Five of these genes, ORFs 6–10 in SaPIbov1, are controlled by LexA (C. Ubeda *et al.*, in preparation) and probably constitute an operon that is provisionally referred to as the encapsidation module. To the left of this module are two or more conserved ORFs of unknown function followed by the replication origin and a gene, *rep,* whose product is similar to helicases and primases. This set of genes is provisionally referred to as the replication module. Further to the left and flanking the point of divergence are two small ORFs with similarity to a variety of regulatory proteins possessing the helix-turn-helix motif and presumably representing a regulatory module (P. Barry and RPN, unpublished data). Still further to the left, in several SaPIs, are two superantigen genes followed by the integrase gene, which is nearly always adjacent to the *att* site. In several of the SaPIs the superantigen and other accessory genes are at the other end, and some have superantigen genes at both ends. Although the overall organization of the SaPI genomes is well conserved, two of them, SaPIm4 and SaPImw2, have diverged widely from the rest as exemplified by the nearest-neighbor tree shown in Figure 10.3. Although ShPI2 from *S. haemolyticus* fits nicely into the SaPI family, SsPI15305 from *S. saprophyticus* is a distant outlier, though it contains all of the SaPI-specific organizational features. SaPI6Δ is at the same site as SaPIn1, SaPIm1, SaPI122 and SaPI2, and SaPI6Δ, has a *ter* gene that is more than 94% identical with the other SaPI *ter* genes, has a degraded *int* gene ψ*int* consisting of a 96 amino acid region that contains a frame shift and is 96% identical to the integrases encoded by the other SaPIs at this site, and has a third ORF (D) that is 74% identical with SaPI1 ORF5, a 70 codon ORF that is widely conserved among

the SaPIs. This element, presumably derived from SaPIs located at this site, is clearly a SaPI remnant, hence the 'Δ'. Nevertheless, it has two other ORFs, A and B, that are not present in any other SaPI or elsewhere in the *S. aureus* genome.

The basic life cycle of a SaPI element (reviewed in Novick, 2003) is initiated by phage-induced excision, using the SaPI-coded integrase. This is presumably by the Campbell mechanism; although the predicted circular monomer can be detected by PCR, it cannot be detected by gel electrophoresis. Like the replication cycle of satellite phage P4, which mimics that of its helper phage P2, it is suggested that the SaPI replication cycle is driven by and mimics that of the helper phage. The known helper phages, with one exception, are generalized transducing phages and therefore use the *pac*-dependent headfull packaging mechanism. Whether they circularize following injection is not known; however, as the linear SaPI genomes injected by phage-like particles do not detectably circularize, it is very likely that replication of both is initiated on linear DNA. It interesting that phage ϕ13, a *cos* phage, can induce excision, circularization, and replication of SaPI1, though it cannot produce infective SaPI1 particles (Ruzin *et al.*, 2001; A. Mathews and RPN, unpublished data). The failure to detect circular SaPI1 molecules following 80α-induced excision suggests that these are rapidly cleaved prior to the initiation of replication. Replication is initiated at a unique origin, characterized by a series of hexa- to octanucleotide repeats, by the Rep protein, which is SaPI-specific (C. Ubeda, *et al.*, unpublished observations) and therefore corresponds to the replicon-specific initiator proteins of other types of replicons (P. Barry, C. Ubeda, and R. P. Novick, unpublished data). Replication is continued by an unidentified replicase, presumably the host replicase, forming a multimeric complex of unknown structure, presumably similar to that of the replicating phage DNA. Several hundred copies of SaPI DNA are produced, along with a similar or possibly smaller number of phage DNA replicas. During this process, the replicating SaPI DNA comigrates with the chromosomal and replicating phage DNAs and probably represents the typical multimeric replicative forms seen with many phages. A SaPI-specific band soon appears in the gel electrophoretic pattern, indistinguishable from that seen immediately following DNA injection by infecting SaPI particles and therefore representing monomeric linear SaPI DNA that has probably been released from mature phage heads. At the completion of the lytic cycle, SaPI DNA is packaged into small-headed phage-like particles, whose head size is about one-third that of the plaque-forming particles, commensurate with the relative sizes of the two genomes, which approach and sometimes surpass the numbers of plaque-forming particles. Packaging

requires a conserved SaPI-coded terminase small subunit, which when complexed with the phage-coded large terminase subunit presumably acts at an unidentified *pac* site, cleaving the multimeric DNA and conducting it to the portal protein that threads the DNA into capsids. It is likely that the SaPI capsids are formed from the same capsomeres as the plaque-forming particles and that SaPI gene(s) determine the size of these. Filling of the heads is followed by *ter*-induced cleavage, generating terminally redundant monomeric SaPI DNA, analogous to that of a typical *pac* phage. Tails are then attached that are indistinguishable from the normal phage tails. The resulting SaPI particles are infectious, giving rise to extremely high transfer frequencies. Following entry, SaPI DNA is site-specifically inserted into its chromosomal *att* site by the Campbell mechanism, utilizing the SaPI integrase. None of the SaPIs identified to date is capable of autonomous replication; the presence of possibly multiple autonomous copies of SaPImw2 (SaPIGlm), identified by PCR in strain mw2 (Baba *et al.*, 2002), almost certainly represents phage-induced excision, since no autonomous SaPI DNA can be detected by gel electrophoresis (A. M. Mathews and R. P. Novick, unpublished data) and since mw2 harbors a prophage capable of inducing excision and replication of SaPIGlm (A. Subedi and R. P. Novick, unpublished data). If the SaPI1 att_S site is occupied by a resident copy, an incoming SaPI1 inserts, at a greatly reduced frequency, into either of the hybrid *att* sites, creating a tandem double, which is unstable and breaks down to yield single insertions of either type, with the excised element being unable to replicate and therefore being lost (A. Subedi and R. P. Novick, unpublished data). Interestingly, the integrases encoded by SaPIbov1 and SaPIbov2 can apparently catalyze spontaneous SaPI excision in the absence of vegetative phage (Ubeda *et al.*, 2003). The excised circular SaPI molecules can be packaged and transduced, in the absence of SaPI replication (Ubeda *et al.*, 2003), by certain phages, presumably as recombinational multimers in normal phage heads. Other SaPI integrases cannot catalyze spontaneous SaPI excision (Lindsay *et al.*, 1998). The basis of this difference is unknown. Further, some SaPIs are induced by phages endogenous to their native host strains, whereas others are not; the basis of phage-SaPI specificity is an area of active investigation.

10.2.5. Resistance Islands

Staphylococci harbor a second unique class of MGEs, namely resistance islands notorious for their possession of the *mecA* gene, which encodes PBP2′, the added penicillin binding protein that is responsible for resistance to methicillin and oxacillin as well as to all other β-lactam antibiotics. The

only other known mobile elements that could be viewed as resistance islands are elements encoding β-lactamase and resistances to cadmium, arsenate, and mercury ions, which are found principally in *agr*-III TSS strains and are genetically linked to SaPI2. These, however, are almost certainly integrated remnants of the well-known penicillinase plasmids, have not been analyzed, and are not considered further here.

The resistance islands encoding PBP2′ are known as SCC*mec* elements, which stands for 'staphylococcal chromosome cassette methicillin'. These range in size from about 20 to more than 60 kb and are always inserted at the same site, near the chromosomal replication origin. They have three defining features in addition to *mecA*, namely a two-subunit site-specific recombinase, *ccrAB,* a unique terminal inverted-direct repeat sequence upon which CcrAB acts, and at least one copy of IS*431* (also known as IS*257*; Figure 10.4; see also color plate after p. 174). CcrAB-catalyzed excision can usually be demonstrated by PCR and is followed either by re-insertion or by loss of the element, which cannot replicate autonomously. The consequence of this is an irreducible, low level of genetic instability. The occurrence of excision/insertion and the presence of identical or nearly identical SCC*mec* elements in otherwise unrelated *S. aureus* strains, and in different staphylococcal species, represents clear if circumstantial evidence for the mobility of these elements; however, this mobility has yet to be demonstrated directly. Although the smaller elements can be moved by generalized transduction, like other chromosomal genes, this is clearly not the general mode of mobility since its frequency is very low, it cannot accommodate the larger elements, and it cannot be responsible for interspecific transfer since staphylococcal phages are highly species specific. DNA-mediated transformation is theoretically possible but is very poorly developed in staphylococci and has not been demonstrated to occur naturally. Far more likely is transfer by plasmid-mediated conjugation. Although SCC*mecs* contain neither the *trs* system nor the classical transfer origin and do not encode relaxation proteins, they always carry IS*431*, an IS element that is very commonly carried by conjugative plasmids, has been implicated in a wide variety of genomic rearrangements (Firth and Skurray, 2006), and could serve as a site of homologous recombination or one-ended transposition between plasmid and SCC*mec*. Additionally, some SCC*mecs* also carry the *pre-RS*$_A$ recombination system (Gennaro *et al.*, 1987), also frequently carried by conjugative plasmids, which could catalyze recombination between plasmid and SCC*mec*. In all of these cases, the putative transfer would be by conduction rather than mobilization, that is, by cointegrate formation and transfer; however, the putative cointegrates have never been seen,

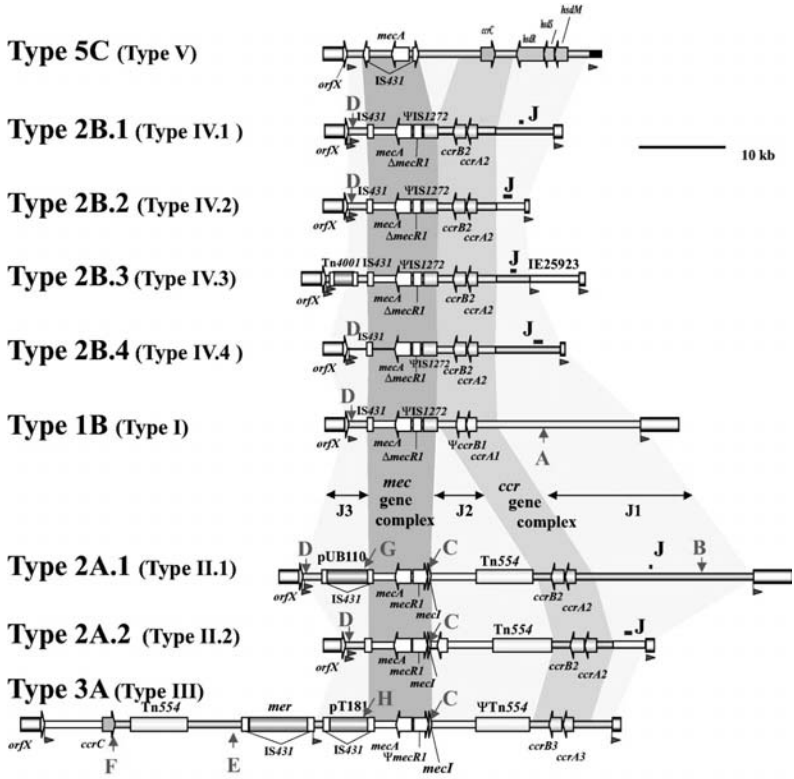


Figure 10.4. Structural comparison of SCC*mec* elements. The structures of SCC*mec* elements are based on the nucleotide sequences deposited in the DDBJ/EMBL/GenBank databases under accession nos. AB033763 (type 1B SCC*mec*), D86934 (type 2A.1 SCC*mec*), AB127982 (type 2A.2), AB037671 (type 3A SCC*mec*), AB063172 (type 2B.1 SCC*mec*), AB063173 (type 2B.2 SCC*mec*), AB096217 (type 2B.3 SCC*mec*), and AB121219 (type 5C). The SCC*mec* element is composed of two essential gene complexes, the *ccr* gene complex (light gray shading) and the *mec* gene complex (medium gray shading). The *ccr* gene complex consists of *ccr* genes that are responsible for the mobility of SCC*mec* and surrounding ORFs. The *mec* gene complex is responsible for methicillin-cephem resistance. Other areas (white) of SCC*mec* are nonessential and are divided into three regions, J1 to J3. Various drug resistance genes are found within the J2 and J3 regions of some SCC*mec* elements. Some ORFs specific for each type are found within the J1 region, such as *pls* in type I SCC*mec* and the *kdp* operon in type II SCC*mec*.
Direct-repeat-containing integration site sequences of SCC*mec* elements are indicated by arrowheads. The locations of primer sets (A to H) used for the multiplex PCR method developed by Oliveira and De Lencastre (2002) are indicated by arrows. The locations of primer sets used for the identification of the J1 region (J) of type 2A and type 2B SCC*mec* element used in this study are indicated by short bars. (Reprinted from Chongtrakool *et al.*, 2006, with the kind permission of the American Society for Microbiology.)

suggesting that if this is the mechanism, it must also involve dissociation following transfer.

In addition to their defining sequence elements, the SCC*mecs* possess a remarkable variety of added genetic units, including transposons, small and large plasmids, and other IS sequences. It is interesting that inserted plasmids are always replication defective, usually because the *rep* gene is insertionally inactivated, and thus they cannot support the replication of excised SCC*mecs*.

A set of typical maps is presented in Figure 10.4. Since all SCC*mecs* are located at the same site, they are therefore all co-ancestral; nevertheless, their occurrence as variants has been important in the epidemiological tracking of *S. aureus* outbreaks and there is consequently a wide-ranging attempt to classify them. At present, there are six different types (R. Daum, personal communication), on the basis of variations in the *ccrAB* and *mecA* genes; however, this process is currently in flux and a consensus scheme is expected in the very near future.

In addition to inserted mobile genetic elements, the SCC*mecs* contain many ORFs with unknown functions, some of which are highly conserved. Although one might imagine that these elements could replicate under the control of, for example, phages, as do the SaPIs, following excision, this has not been seen and thus the unknown ORFs are not likely to represent replication functions.

10.3. EVOLUTION AND DISSEMINATION OF THE STAPHYLOCOCCAL MGEs AND THEIR GENES

It is generally assumed that MGEs such as plasmids evolved subsequently to the initial evolution of the cellular pattern of biological organization – that is, a replication system confined within a semi-permeable enveloping membrane. The only fossil record, however, is extant genomes, whose sequences do not support even a speculative interpretation of the earliest events. Therefore, it is entirely possible that the MGEs evolved concomitantly with the prokaryotic cell. For example, if one assumes that the appearance of self-replicating nucleic acids predated their membranous confinement, it is perfectly possible that different types of replicons may have evolved independently in the pre-cellular biosphere and that two or more may have been enclosed in some of the earliest prokaryotic cells. This concept is indirectly supported by the inference that MGEs are independently evolving units ('selfish DNA') that use host cells as their environmental niches (Novick, 1980).

One must envision a rather different history for transposons, which obviously could have evolved only within the confines of pre-existing genetic material. The remarkably simple basic plan of a transposon consists of a terminally repeated sequence acted upon by a transposase – a recombinase that catalyzes exchange between one fixed and one more or less random partner. Though it is idle to speculate on the evolutionary history of this simple plan, it has probably evolved many times and its spectacular record of evolutionary success despite the usual lack of any selective advantage for the organism has reinforced the concept of selfish DNA. Transposons have had a major role in genomic plasticity throughout the phylogenetic spectrum, but a downside to their promiscuity is that they not only cause genetic diseases, including cancer, but also may gradually swamp the genome, with highly deleterious consequences for life as we know it. Although transposons do not seem to have had a major role in the evolution of phage genomes – perhaps because these have developed their own very effective means of genetic exchange – they have had a major impact on the evolution of plasmids and chromosomal islands. Having incorporated particular genes, especially antibiotic resistance genes, they have contributed mightily to the spread of these genes, through their own agency as well as that of the plasmids to which they have attached.

With respect to plasmids, given a primordial replicon, one may sketch a credible scenario for subsequent events leading to extant genomes. For most extant plasmids, the essential elements of the replicon, consisting of a replication origin, an initiator gene, and a copy control system, are localized – that is, are not interrupted by non-essential genetic units (Novick, 1967; Timmis *et al.*, 1975). And there are a few very small plasmids that carry no other gene functions. These would have evolved further by acquiring non-essential genes that could affect the host cell's phenotype. Obviously, insertion of new DNA occurred at a non-essential site in the replicon and must have occurred by non-homologous recombination. In many cases, transposons were clearly involved; alternatively, a variety of enzymes, such as restriction-like enzymes, could catalyze the insertion of new DNA. Once a new gene is inserted, it can serve as a non-essential site for subsequent insertions. There are several interesting questions with respect to the evolution of extant genetic configurations: (1) What are the advantages and disadvantages to either the host or the MGE of the presence of a given gene on an MGE? (2) What are the evolutionary pathways by which MGEs acquire genes that affect the host cell's phenotype? (3) What are the metabolic consequences of MGEs for the host organism? Although much of the thinking with respect to these questions is

related to plasmids, they apply in one way or another to the other MGEs as well.

10.3.1. Advantages and/or Disadvantages

Needless to say, antibiotic resistance genes represent a major selective advantage, though it is not always obvious whether plasmid linkage is advantageous to the organism. Erythromycin resistance conferred by erythromycin ribosome methylase (Erm) versus chloramphenicol resistance (chloramphenicol transacetylase, Cat) offers an interesting example. Erm provides absolute resistance in single copy, is therefore equally effective whether on a plasmid or in the chromosome (via transposon), and frequently occurs in either location. Staphylococcal Cat confers very weak resistance in single copy and is found only on high-copy plasmids. *E. coli* Cat is effective in single copy and is found either on plasmids or on a (chromosomal) transposon. A similar situation exists with tetracycline resistance, TcR, which occurs in many forms. Two of these, *tetM* and *tetA(K)*, are distinguished by their relative efficacy – *tetM* provides full resistance in single copy and is often found on the chromosome, carried by at least one known transposon, whereas *tetA(K)* confers very weak resistance in single copy and is found only on high-copy plasmids. Similar arguments can be made for other resistance determinants, from which it can be suggested that individual features of the resistance gene are important determinants of their preferred locations. Regarding the importance of horizontal transfer in the carriage of resistance genes by MGEs, it is suggested that horizontal transfer is more advantageous to an MGE, because it promotes spread of the element, than to the host, because it enables the survival of competing organisms. Resistance genes carried by MGEs, especially plasmids, are additionally advantageous to the MGE as well as to the host organism, since they offset the metabolic cost of the MGE to a newly infected host by their selective advantage in a sea of antibiotics.

Since transposons may bring about the transfer of any chromosomal gene to a plasmid, the question arises of why it is so rare to find standard or essential chromosomal genes on plasmids. One possible answer is that since such genes function well in single copy, there is no advantage to plasmid carriage, and since acquisition of such a gene will only duplicate what is already present, it will not offset the metabolic cost of a new plasmid and will not confer any evolutionary advantage for either plasmid or host. An interesting exception to this rule is the plasmid carriage of dihydrofolate reductase, conferring resistance to sulfonamides. Although *dhfR* approximates an

essential chromosomal gene, its acquisition may not represent a simple duplication and may be highly selective: certain potential donor organisms possess a variety of the enzyme that is naturally much more resistant to inhibition by sulfonamides than the endogenous counterpart of a recipient organism such as *E. coli*; the acquisition of such a gene on a plasmid has thus resulted in classical sulfonamide resistance in *E. coli*.

10.3.2. Pathways of Acquisition

It is very likely that any gene carried by an MGE originated in some other organism and was horizontally transferred. A well-established scenario is that any antibiotic-producing bacterium must be resistant to its own toxic product and has therefore evolved resistance along with the evolution of the antibiotic. Conjugative plasmids and transposons are common in antibiotic producers such as *Streptomyces* sp., and any such element could acquire an endogenous resistance gene, by several possible mechanisms, and then transfer it to other species, including human and animal pathogens.

10.3.3. Stability and Persistence

Plasmids clearly represent a metabolic cost in proportion to their sizes, copy numbers, and the levels of expression of their genes. This has given rise to the general concept that plasmids are subject to ejection unless they provide some tangible benefit to their host. This concept is supported by fitness studies in which a strain into which a plasmid has been placed is outcompeted by the original strain in a mixed culture. There are two problems with this type of experiment – first, growth in a flask may have little or nothing to do with fitness in a natural setting, and second, the experimental paradigm is flawed by the choice of strains. When a naturally occurring plasmid-carrying strain is competed against the same strain that has been cured of its plasmid, no such difference in fitness is seen. This is because the naturally occurring strain has evolved to maintain the plasmid, has adapted its metabolism to the plasmid's presence, and is not actually paying any metabolic tax as a consequence. This argument is underscored by a number of facts: Many plasmids are cryptic – do not contain any gene that affects the host's phenotype – yet are perfectly stable; in some bacteria, plasmids make up a substantial proportion of the cell's genetic content – perhaps 20–30% in strains of *Bacillus megaterium* or *Borrelia burgdorferi*, and many of these are cryptic; and finally, plasmids are selfish in that they autonomously regulate their own replication and have elaborate means of ensuring their survival,

including rapid copy number correction, partitioning mechanisms, and, in certain cases, the destruction of host cells that have failed to transmit them during cell division. Nevertheless, plasmids do have finite loss rates and it has been suggested that their survival depends on frequencies of horizontal transfer that must be sufficient to balance the loss rates. Although the logic here is unassailable, it depends on equivalence of fitness; for example, the plasmid-carrier may actually be more fit in the natural environment than the plasmid-loser, even if the plasmid is cryptic. In such a case, the logic breaks down.

Prophages are also generally very stable; they minimize the metabolic costs of gene expression by keeping their genome shut down in the prophage state. Occasional excision would theoretically lead to loss since the genome would remain shut down and unable to replicate. This is countered by the invariable presence of active phage particles, released by spontaneous induction, which would soon re-infect any cell that happened to lose its copy.

Transposons, on the other hand, are not only extremely stable whether or not they represent any metabolic burden, but also have the ability to spread within a host; consequently, they are rarely if ever lost under natural conditions. On the contrary, they pose a threat to take over plant and mammalian genomes, representing the most aggressive brand of selfish DNA yet to be identified.

REFERENCES

Arnold, H. P., She, Q., Phan, H., *et al.* (1999). The genetic element pSSVx of the extremely thermophilic crenarchaeon *Sulfolobus* is a hybrid between a plasmid and a virus. *Mol Microbiol*, **34**, 217–26.

Baba, T., Takeuchi, F., Kuroda, M., *et al.* (2002). Genome and virulence determinants of high virulence community-acquired MRSA. *Lancet*, **359**, 1819–27.

Bensing, B. A., Lopez, J. A., and Sullam, P. M. (2004). The *Streptococcus gordonii* surface proteins GspB and Hsa mediate binding to sialylated carbohydrate epitopes on the platelet membrane glycoprotein Ibalpha. *Infect Immun*, **72**, 6528–37.

Berg, T., Firth, N., Apisiridej, S., *et al.* (1998). Complete nucleotide sequence of pSK41: evolution of staphylococcal conjugative multiresistance plasmids. *J Bacteriol*, **180**, 4350–9.

Birch, P., and Khan, S. A. (1992). Replication of single-stranded plasmid pT181 DNA in vitro. *Proc Natl Acad Sci USA*, **89**, 290–4.

Chatterjee, A. N. (1969). Use of bacteriophage-resistant mutants to study the nature of the bacteriophage receptor site of *Staphylococcus aureus*. *J Bacteriol*, **98**, 519–27.

Chongtrakool, P., Ito, T., Ma, X. X., *et al.* (2006). Staphylococcal cassette chromosome mec (SCC*mec*) typing of methicillin-resistant *Staphylococcus aureus* strains isolated in 11 Asian countries: a proposal for a new nomenclature for SCC*mec* elements. *Antimicrob Agents Chemother*, **50**, 1001–12.

Coleman, D. C., Sullivan, D. J., Russell, R. J., *et al.* (1989). *Staphylococcus aureus* bacteriophages mediating the simultaneous lysogenic conversion of β-lysin, staphylokinase and enterotoxin A: molecular mechanism of triple conversion. *J Gen Microbiol*, **135**, 1679–97.

Diep, B. A., Gill, S. R., Chang, R. F., *et al.* (2006). Complete genome sequence of USA300, an epidemic community-acquired methicillin-resistant *Staphylococcus aureus* strain. *Lancet*, **367**, 731–9.

Firth, N., and Skurray, R. A. (1998). Mobile elements in the evolution and spread of multiple-drug resistance in staphylococci. *Drug Resist Updates*, **1**, 49–58.

Firth, N., and Skurray, R. A. (2006). Genetics: accessory elements and genetic exchange. In Fischetti, V. A., Novick, R. P., Ferretti, J. J., Portnoy, D. A., and Rood, J. I. (Eds.). *Gram-positive pathogens* (2nd ed.). Washington, DC: ASM Press.

Firth, N., Apisiridej, S., Berg, T., *et al.* (2000). Replication of staphylococcal multiresistance plasmids. *J Bacteriol*, **182**, 2170–8.

Fitzgerald, J. R., Monday, S. R., Foster, T. J., *et al.* (2001). Characterization of a putative pathogenicity island from bovine *Staphylococcus* aureus encoding multiple superantigens. *J Bacteriol*, **183**, 63–70.

Gennaro, M. L., Kornblum, J., and Novick, R. P. (1987). A site-specific recombination function in *Staphylococcus aureus* plasmids. *J Bacteriol*, **169**, 2601–10.

Gill, S. R., Fouts, D. E., Archer, G. L., *et al.* (2005). Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant *Staphylococcus aureus* strain and a biofilm-producing methicillin-resistant *Staphylococcus epidermidis* strain. *J Bacteriol*, **187**, 2426–38.

Gillet, Y., Issartel, B., Vanhems, P., *et al.* (2002). Association between *Staphylococcus aureus* strains carrying gene for Panton-Valentine leukocidin and highly lethal necrotising pneumonia in young immunocompetent patients. *Lancet*, **359**, 753–9.

Groisman, E. A., and Ochman, H. (1996). Pathogenicity islands: bacterial evolution in quantum leaps. *Cell*, **87**, 791–4.

Gruss, A. D., Ross, H. F., and Novick, R. P. (1987). Functional analysis of a palindromic sequence required for normal replication of several staphylococcal plasmids. *Proc Natl Acad Sci USA*, **84**, 2165–9.

Highlander, S. K., and Novick, R. P. (1990). Mutational and physiological analyses of plasmid pT181 functions expressing incompatibility. *Plasmid*, **23**, 1–15.

Hochhut, B., Dobrindt, U., and Hacker, J. (2005). Pathogenicity islands and their role in bacterial virulence and survival. *Contrib Microbiol*, **12**, 234–54.

Holden, M. T., Feil, E. J., Lindsay, J. A., *et al.* (2004). Complete genomes of two clinical *Staphylococcus aureus* strains: evidence for the rapid evolution of virulence and drug resistance. *Proc Natl Acad Sci USA*, **101**, 9786–91.

Johnson, L. B., Saeed, S., Pawlak, J., Manzor, O., and Saravolatz, L. D. (2006). Clinical and laboratory features of community-associated methicillin-resistant *Staphylococcus aureus*: is it really new? *Infect Control Hosp Epidemiol*, **27**, 133–8.

Khan, S. A. (2005). Plasmid rolling-circle replication: highlights of two decades of research. *Plasmid*, **53**, 126–36.

King, M. D., Humphrey, B. J., Wang, Y. F., *et al.* (2006). Emergence of community-acquired methicillin-resistant *Staphylococcus aureus* USA 300 clone as the predominant cause of skin and soft-tissue infections. *Ann Intern Med*, **144**, 309–17.

Kramer, M. G., Espinosa, M., Misra, T. K., and Khan, S. A. (1998). Lagging strand replication of rolling-circle plasmids: specific recognition of the *ssoA*-type origins in different gram-positive bacteria. *Proc Natl Acad Sci USA*, **95**, 10505–10.

Krolewski, J. J., Murphy, E., Novick, R. P., and Rush, M. G. (1981). Site-specificity of the chromosomal insertion of *Staphylococcus aureus* transposon Tn*554*. *J Mol Biol*, **152**, 19–33.

Kuroda, M., Ohta, T., Uchiyama, I., *et al.* (2001). Whole genome sequencing of methicillin-resistant *Staphylococcus aureus*. *Lancet*, **357**, 1225–40.

Kuroda, M., Yamashita, A., Hirakawa, H., *et al.* (2005). Whole genome sequence of *Staphylococcus saprophyticus* reveals the pathogenesis of uncomplicated urinary tract infection. *Proc Natl Acad Sci USA*, **102**, 13272–7.

Kwan, T., Liu, J., Dubow, M., Gros, P., and Pelletier, J. (2005). The complete genomes and proteomes of 27 *Staphylococcus aureus* bacteriophages. *Proc Natl Acad Sci USA*, **102**, 5174–9.

Lee, C. Y., and Iandolo, J. J. (1986). Lysogenic conversion of staphylococcal lipase caused by insertion of the bacteriophage L54a genome into the lipase structural gene. *J Bacteriol*, **166**, 385–91.

Lina, G., Piemont, Y., Godail-Gamot, F., *et al.* (1999). Involvement of Panton-Valentine leukocidin-producing *Staphylococcus aureus* in primary skin infections and pneumonia. *Clin Infect Dis*, **29**, 1128–32.

Lindqvist, B. H., Deho, G., and Calendar, R. (1993). Mechanisms of genome propagation and helper exploitation by satellite phage P4. *Microbiol Rev*, **57**, 683–702.

Lindsay, J. A., Ruzin, A., Ross, H. F., Kurepina, N., and Novick, R. P. (1998). The gene for toxic shock toxin is carried by a family of mobile pathogenicity islands in *Staphylococcus aureus*. *Mol Microbiol*, **29**, 527–43.

Lyon, B. R., and Skurray, R. (1987). Antimicrobial resistance of *Staphylococcus aureus*: genetic basis. *Microbiol Rev*, **51**, 88–134.

Macrina, F. L., and Archer, G. L. (1993). Conjugation and broad host range plasmids in streptococci and staphylococci. In Clewell, D. B. (Ed.). *Bacterial conjugation*. New York: Plenum.

Marrack, P., Winslow, G. M., Choi, Y., *et al.* (1993). The bacterial and mouse mammary tumor virus superantigens; two different families of proteins with the same functions. *Immunol Rev*, **131**, 79–92.

Musser, J. M., Schlievert, P. M., Chow, A. W., *et al.* (1990). A single clone of *Staphylococcus aureus* causes the majority of cases of toxic shock syndrome. *Proc Natl Acad Sci USA*, **87**, 225–9.

Nakayama, J., Igarashi, S., Nagasawa, H., *et al.* (1996). Isolation and structure of staph-cAM373 produced by *Staphylococcus aureus* that induces conjugal transfer of *Enterococcus faecalis* plasmid pAM373. *Biosci Biotechnol Biochem*, **60**, 1038–9.

Narita, S., Kaneko, J., Chiba, J., *et al.* (2001). Phage conversion of Panton-Valentine leukocidin in *Staphylococcus aureus*: molecular analysis of a PVL-converting phage, phiSLT. *Gene*, **268**, 195–206.

Nimmo, G. R., Coombs, G. W., Pearson, J. C., *et al.* (2006). Methicillin-resistant *Staphylococcus aureus* in the Australian community: an evolving epidemic. *Med J Aust*, **184**, 384–8.

Novick, R. (1976). Plasmid-protein relaxation complexes in *Staphylococcus aureus*. *J Bacteriol*, **127**, 1177–87.

Novick, R. P. (1967). Mutations affecting replication and maintenance of penicillinase plasmids in *S. aureus*. Paper presented at the Fifth International Congress of Chemotherapy, Vienna.

Novick, R. P. (1980). Plasmids. *Sci Am*, **243**, 102–4, 6, 10 passim.

Novick, R. P. (1987). Plasmid incompatibility. *Microbiol Rev*, **51**, 381–95.

Novick, R. P. (2003). Mobile genetic elements and bacterial toxinoses: the superantigen-encoding pathogenicity islands of *Staphylococcus aureus*. *Plasmid*, **49**, 93–105.

Novick, R. P., Iordanescu, S., Projan, S. J., Kornblum, J., and Edelman, I. (1989). pT181 plasmid replication is regulated by a countertranscript-driven transcriptional attenuator. *Cell*, **59**, 395–404.

Novick, R. P., Schlievert, P., and Ruzin, A. (2001). Pathogenicity and resistance islands of staphylococci. *Microb Infect*, **3**, 585–94.

Oliveira, D. C., and De Lencastre, H. (2002). Multiplex PCR strategy for rapid identification of structural types and variants of the *mec* element in methicillin-resistant *Staphylococcus aureus*. *Antimicrob Agents Chemother*, **46**, 2155–61.

Paulsen, I. T., Firth, N., and Skurray, R. A. (1997). Resistance to antimicrobial agents other than β-lactams. In Crossley, K. B., and Archer, G. L. (Eds.). *The staphylococci in human disease*. New York: Churchill Livingstone.

Projan, S. J., and Archer, G. L. (1989). Mobilization of the relaxable *Staphylococcus aureus* plasmid pC221 by the conjugative plasmid pG01 involves three pC221 loci. *J Bacteriol*, **171**, 1841–5.

Rasooly, A., and Novick, R. P. (1993). Replication-specific inactivation of the pT181 plasmid initiator protein. *Science*, **262**, 1048–50.

Ruzin, A., Lindsay, J., and Novick, R. P. (2001). Molecular genetics of SaPI1 – a mobile pathogenicity island in *Staphylococcus aureus*. *Mol Microbiol*, **41**, 365–77.

Stout, V. G., and Iandolo, J. J. (1990). Chromosomal gene transfer during conjugation by *Staphylococcus aureus* is mediated by transposon-facilitated mobilization. *J Bacteriol*, **172**, 6148–50.

Takeuchi, F., Watanabe, S., Baba, T., *et al*. (2005). Whole-genome sequencing of *Staphylococcus haemolyticus* uncovers the extreme plasticity of its genome and the evolution of human-colonizing staphylococcal species. *J Bacteriol*, **187**, 7292–308.

Thomas, W. D., Jr., and Archer, G. L. (1989). Identification and cloning of the conjugative transfer region of the *S. aureus* plasmid pG01. *J Bacteriol*, **171**, 684–91.

Timmis, K., Cabello, F., and Cohen, S. N. (1975). Cloning, isolation, and characterization of replication regions of complex plasmid genomes. *Proc Natl Acad Sci USA*, **72**, 2242–6.

Ubeda, C., Tormo, M. A., Cucarella, C., *et al*. (2003). Sip, an integrase protein with excision, circularization and integration activities, defines a new family of mobile *Staphylococcus aureus* pathogenicity islands. *Mol Microbiol*, **49**, 193–210.

Udo, E. E., and Jacob, L. E. (1998). Conjugative transfer of high-level mupirocin resistance and the mobilization of non-conjugative plasmids in *Staphylococcus aureus*. *Microb Drug Resist*, **4**, 185–93.

Waldor, M. K., and Friedman, D. I. (2005). Phage regulatory circuits and virulence gene expression. *Curr Opin Microbiol*, **8**, 459–65.

Weigel, L. M., Clewell, D. B., Gill, S. R., *et al*. (2003). Genetic analysis of a high-level vancomycin-resistant isolate of *Staphylococcus aureus*. *Science*, **302**, 1569–71.

Winkler, K. C., Wart, J. D., and Grootsen, C. (1965). Lysogenic conversion of staphylococci to loss of β-toxin. *J Gen Microbiol*, **39**, 321–33.

Wu, S. W., De Lencastre, H., and Tomasz, A. (1999). The *Staphylococcus aureus* transposon Tn551: complete nucleotide sequence and transcriptional analysis of the expression of the erythromycin resistance gene. *Microb Drug Resist*, **5**, 1–7.

Zhang, X., McDaniel, A. D., Wolf, L. E., *et al.* (2000). Quinolone antibiotics induce Shiga toxin-encoding bacteriophages, toxin production, and death in mice. *J Infect Dis*, **181**, 664–70.

Zhang, Y. Q., Ren, S. X., Li, H. L., *et al.* (2003). Genome-based analysis of virulence genes in a non-biofilm-forming *Staphylococcus epidermidis* strain (ATCC 12228). *Mol Microbiol*, **49**, 1577–93.

Zou, D., Kaneko, J., Narita, S., and Kamio, Y. (2000). Prophage, phiPV83-pro, carrying Panton-Valentine leukocidin genes, on the *Staphylococcus aureus* P83 chromosome: comparative analysis of the genome structures of phiPV83-pro, phiPVL, phi11, and other phages. *Biosci Biotechnol Biochem*, **64**, 2631–43.

CHAPTER 11

Influence of Human Lifestyle on Dissemination of Transferable Elements

Wolfgang Witte

11.1. INTRODUCTION

During the past 20,000 years the most striking change in the lifestyle of humans was the transition from the hunter-gatherer culture to the introduction of agriculture and animal husbandry associated with stable settlement in the Neolithic. With sufficient food resources, the human population, which had been scattered, started to grow. Larger urban communities were formed, which was the starting point of culture and technology.

Humans have evolved with bacterial communities (microecological systems) as colonizers (skin, mucosa, gastrointestinal tract) and as conditional pathogens. For true pathogens the more dense human communities became of particular interest, especially for specialization in the human hosts (McKeown, 1988).

The so-called technical revolution that began at the end of the 19th century created a need for energy as a permanent concern. Because of continuing urbanization and the growing human population in the Third World, sustainability of food supply remains an important issue.

When, 200 years ago, a more productive agriculture began that was based on the then-young agricultural sciences, the principle of enduring means of production long endured, especially with regard to recycling of energy. These principles began to fade with the invention of mineral fertilizers, followed by mechanization and energy-consuming (wasting) means of production that were more independent of seasons. By the middle of the last century this had led to a high degree of mechanization of agriculture, where animals and plants were regarded more as "work pieces" than as living beings. The earlier and basically rational principle of enduring means of production was no longer followed. This became most pertinent with mass animal production

where thousands of animals were raised unnaturally in dense communities. This situation became rather attractive for bacterial pathogens.

Frequent meat consumption, earlier a privilege of the comparably small wealthy part of the human society, became a nutritional habit at least in the industrialized parts of the world.

Two conditions favor dissemination of bacterial pathogens among animals in mass production: the dense population and the general stress on the animals that are due to the artificial conditions under which they are raised. The regular and permanent use of antibiotics as feed additives was a consequence. The real bases of the mode of action of the so-called growth promoters had never been really explained; most likely growth promoters have a prophylactic effect with respect to infectious diseases. Growth promoters used in countries of the European communities, until their ban by the end of 2002, are listed in Table 11.1. The ban followed the precautionary principle and was based on mounting evidence for the creation of a non-calculable risk of a large reservoir of transferable antibiotic resistance created by permanent selective pressure imposed by growth promoters (for references, see Witte *et al.*, 2002).

In this chapter, the influence of human lifestyle on the spread of transferable elements is described more generally with respect to changes that began in the Neolithic and in particular with respect to growth promoters and animal husbandry.

When possible, comparison of genomic sequences of bacterial species restricted to particular hosts or ecological niches with those of free-living relatives or ancestors reveals that various changes occur soon after new ecological niches are invaded (Moran and Plague, 2004). In particular a rapid increase in frequencies of mobile elements is associated with numerous genomic rearrangements and deletions (Moran and Plague, 2004; Parkhill *et al.*, 2001). For the following bacterial species, which are pathogens of domesticated plants and of livestock, these processes are very likely the consequence of the substantial changes in human lifestyle that began in the Neolithic.

Burkholderia mallei as an obligate parasite of horses and related species such as donkeys and their hybrid offspring: in comparison to *Burkholderia pseudomallei*, its genome contains an extraordinarily higher number of insertion sequence (IS) elements (Nierman *et al.*, 2004). The domestication of horses and consequently a considerable increase in the horse population began about 4,500 years ago (Jansen *et al.*, 2002).

Another example is represented by *Xanthomonas* spp. In contrast to other species of this genus, the genome of the rice pathogen *Xanthomonas oryzae* contains the high number of more than 800 IS elements; this is likely

associated with the specialization of *X. oryzae* in rice plants after intense culturing of rice across South-East Asia (Da Silva *et al.*, 2002).

A high load of IS elements also exists among pathogenic enteric bacteria, in particular in *Shigella* spp. and *S. enterica* serovar Typhi if compared to close relatives with less host restriction (Jin *et al.*, 2002; Wei *et al.*, 2003). Integration and excision of IS elements can lead to pseudogenes. The time scale of IS expansion had been calculated by looking at the divergence between IS-interrupted open reading frames (ORFs) and functionally homologous ones in related bacterial strains; although these estimates are fairly rough, the degrees of nucleotide divergence for the IS-generated pseudogenes are close to the Neolithic frame (Mira *et al.*, 2006).

Altogether, sequence comparisons of prominent members of five IS families reveals low sequence divergence with respect to other parts of the genomes. In long evolutionary terms, the increase of IS in bacterial genomes is probably detrimental. Therefore, IS may undergo periodic extinction in bacterial lineages (Wagner, 2006).

Antibiotics have been used as growth promoters in animal feeding for several decades. Soon after the detection of transferable antibiotic resistance in Enterobacteriaceae, animal production was identified as a potential reservoir of resistance determinants. At that time, oxytetracycline was the main antibiotic used as a feed additive, and a number of investigations showed a strong association between ergotropic tetracycline use and the frequency of tetracycline resistance in enterobacterial isolates from animals and livestock workers (Levy *et al.*, 1976a,b).

As early as 1969, a report to the British government recommended that antibiotics that are used in human therapy or antibacterial substances that select for resistance against themselves should not be used as growth promoters in animal feeding. This demand was renewed by the World Health Organization (WHO) in 1994 and again in 1997.

In the 1980s, relevant EC authorities established legislation for the use of antibacterial agents as growth promoters in animal husbandry (EU Guideline 87/163), prohibiting the use of substances selecting for resistance to antibacterial agents used in human therapy (Helmuth and Bulling, 1985).

Although new candidate antibacterial growth promoters should have fulfilled this criterion to be passed by regulatory boards in EC countries, there are a number of antibiotics that had been used as growth promoters for more than 20 years, and for which cross-resistance to antibiotics used in human therapy is known (Table 11.1). From the end of the 1960s to the end of the 1990s, there was an ongoing debate over whether and to what extent antibiotic use as feed additives contributes to resistance development in human

Table 11.1. *Antibacterial substances used as feed additives in animal husbandry in European countries*

Compound	Class	Mode of action	Known resistance mechanisms	Cross-resistance
Olaquindox	quinoxaline	inhibition of DNA synthesis	no reduction	other quinoxalines
Zn-Bacitracin	polypeptide	inhibits dephosphorylation of undecaprenyl pyrophosphate lipid carrier of cell wall precursors	spontaneous mutant, not characterized	unknown
Flavomycin	phosphoglycolipid	inhibition of cell wall synthesis	unknown	unknown
Monensin, salinomycin	ionophore	disaggregation of cytoplasmic membrane	unknown	unknown
Tylosin	macrolide	binding to the 60S ribosomal subunit inhibition of binding aminoacyl-tRNAs and of peptidyl transferase	methylation of 23S rRNA in position 2058	when constitutive; all macrolides, lincosamides, streptogramin B-antibiotics
Virginiamycin	streptogramin	inhibition of conformational change of the 24S ribosomal protein, preventing protein chain extension	O-hydrolase, porter mechanism, acetyltransferases	other streptogramin antibiotics
Avoparcin	glycopeptide	inhibition of transglycosylation	modification of the peptidoglycan precursor	vancomycin, teicoplanin, daptomycin, avoparcin
Avilamycin	oligosaccharide	inhibition of formylmethionine binding to ribosome	methylation of 23S rRNA in position 2470	everninomycin

bacterial pathogens. In the end, the use of antimicrobials as growth promoters was phased out between 2000 and 2006 in countries of the European community.

11.2. MOLECULAR EPIDEMIOLOGY

Looking at the ergotropic use of these antibiotics as one of the possible sources for emergence and spread of transferable resistance, the following questions have to be answered:

1. Does ergotropic use of antibiotics really select for resistance?
2. Is cross-resistance to antibiotics used in human therapy determined by the same resistance genes in strains from both animal and human sources?
3. Are the corresponding resistance determinants disseminated via the food chain, and to what extent are they found in the intestinal flora of non-hospitalized humans?

Once a resistance gene has become widely disseminated among the community, it is always difficult to trace it back to its origin. Circumstantial evidence can be provided by molecular subtyping of the resistance determinant(s) or by a prospective study starting with the introduction of use of a particular antibiotic as growth promoters.

11.2.1. Streptothricin Resistance in *E. coli*

A prospective study became possible when, in the former Eastern Germany, the streptothricin antibiotic nourseothricin replaced oxytetracycline in animal feeding in 1983 and was used nationwide for animal feeding only. No resistance was seen in Enterobacteriaceae from animals and humans at that time. The first occurrence of a transposon-coded resistance mechanism (streptothricin acetyltransferase) was observed 2 years later in *Escherichia coli* from the gut flora of pigs. By the time its use was stopped after German reunification in 1990, resistance had spread to *E. coli* from the gut flora of pig farmers and their family members, and also to *E. coli* from citizens of municipal communities and from urinary tract infections. The resistance determinants obviously spread in the absence of any selective pressure. Finally, the resistance determinant was detected in *Salmonella* and *Shigella* from cases of diarrhea (Hummel *et al.*, 1986; Tschäpe, 1994).

11.2.2. Glycopeptide Resistance in Enterococci

The emergence of glycopeptide resistance in *Enterococcus faecium* (GREF) already resistant to other antibiotics in nosocomial infections and the demonstration of GREF in sewage treatment plants of towns without hospitals had again focused attention on the use of the glycopeptide avoparcin as a growth promoter.

Although the mechanisms of glycopeptide resistance in enterococci are phenotypically and genotypically heterogeneous, they are based on the same principle; cleavage of the terminal D-alanine from the D-ala-D-ala end-groups of muraminic acid as cell wall precursor by peptidases and replacement by a D-β-hydroxylic acid due to a ligase activity. Glycopeptides are unable to block the corresponding depsipeptides as they do with D-ala-D-ala muraminic acid. The genes of the actual resistance mechanism, namely ligase, peptidase, and dehydrogenase, are inactive in the absence of glycopeptides. Their expression is regulated by two components: a sensor (membrane protein with histidine kinase activity) is activated by disturbance in cell wall synthesis due to glycopeptides (or certain other cell wall active antibiotics) and dephosphorylates a response regulator, which activates the promoter region of the resistance operon (Arthur and Courvalin, 1993).

In case of the most frequently observed VanA genotype, which mediates high-level glycopeptide resistance, vancomycin, teicoplanin, and avoparcin are able to induce its expression. In case of the VanB genotype and the recently described VanD genotype, only vancomycin can induce expression. Genotypes C_1, C_2, and C_3 code for low-level constitutive vancomycin resistance as a natural (intrinsic) resistance of the species *E. gallinarum, E. casseliflavus*, and *E. flavescens* (without clinical significance).

The *vanA* gene cluster is localized on transposons of the Tn*1546* type (Arthur *et al.*, 1996). In most cases these transposons are integrated into conjugative plasmids conferring conjugation among enterococci but also from enterococci to other Gram-positive bacteria.

Aimed studies in animal flocks with and without avoparcin use have confirmed that avoparcin has a significant role in selecting for emergence and spread of GREF in animal husbandry (Klare, 1995a; Table 11.2), as did a large study in Denmark (Aarestrup, 1995).

Dissemination of GREF via meat products: If GREF are found in the intestinal flora of meat animals, their detection in all of 12 investigated broiler chicken carcasses and in low numbers in five of 13 samples of minced pork from different producers was not a surprise (Klare *et al.*, 1993). In a Dutch study GREF were detected in stool samples from meat eaters but not in samples from strict vegetarians (Schouten *et al.*, 1997).

Table 11.2. *Glycopeptide-resistant* Enterococcus faecium *outside hospitals*

No.	Location (German states)	Use of avoparcin	Detection of VRE liquid manure
2	pig farm (Sachsen-Anhalt, Thüringen)	+	+
1	chicken farm (Sachsen-Anhalt)	+	+
1	egg laying hens, farm (Sachsen-Anhalt)	−	−
8	bio farms (Baden-Württemberg)	−	−
13	individual pig holdings (Sachsen-Anhalt)	−	−
11	individual egg laying hens (Sachsen-Anhalt)	−	−

VRE, vancomycin-resistant *Enterococcus*.

The use of avoparcin was ended by national legislation in Denmark in 1995, in Germany in January 1996, and in all EC countries in April 1997 by EC-wide legislation. As a result, by the end of 1997 GREF were found in only 25% of investigated broiler chicken and only in 3.3% of non-hospitalized carriers ($n = 400$; Klare *et al.*, 1999).

Antibiotic resistance can be disseminated with particular bacterial strain clones or by horizontal transfer among different strains. We found the *vanA* gene in *E. faecium* strains of different ecological origin (meat animals, meat products, sewage, healthy persons, and nosocomial infections) exhibiting a variety of different *Sma*I macro-restriction patterns; GREF are polyclonal (Klare *et al.*, 1995b). In 21 investigated strains the *vanA* gene cluster was found integrated into different conjugative plasmids and a frequent in vitro transfer among different *E. faecium* strains. A polymorphism due to insertions into the non-coding regions of the *vanA* gene cluster has been described for GREF from the United States and from western European countries (Miele *et al.*, 1995). This polymorphism allows molecular typing of the *vanA* gene cluster. The occurrence of the same subtypes in GREF of human and animal origin indicates that the resistance gene pools in GREF of human and of animal origin communicate (Table 11.3).

11.2.3. Streptogramin Resistance in Enterococci

A situation parallel to avoparcin resistance is observed for streptogramins. In case of infections with GREF the quinupristin/dalfopristin

Table 11.3. *Molecular characterization of the vanA-gene cluster of* Enterococcus faecium *from human and animal sources**

Origin	Methodology applied	Results	References
D	overlapping PCR	majority of isolates from humans, poultry, and pig uniform	Werner *et al.*, 1997
DK	overlapping PCR, sequencing of *vanX*	type 1: humans, chicken, turkey type 2: humans, pigs type 3: humans type 4: humans types 5–16: rare types of various sources	Jensen *et al.*, 1998
GB	overlapping PCR (10 primer pairs)	24 different types type A: 34% of non-humans sources type H: 32% of human, 5% non-human types T, U, W: 4–10% from human and non-human sources	Woodford *et al.*, 1998
N	RFLP of long PCR	uniform for isolates from humans and from chicken	Simonsen *et al.*, 1998
NL	overlapping PCR	*vanX-vanY*: 13,300 bp: poultry (42%) 543 bp: poultry (59%) 543 bp: humans (100%)	Van den Braak *et al.*, 1998

* For more details, see Section 11.2.2.

PCR, polymerase chain reaction; RFLP, restriction fragment length polymorphism.

combination Synercid is one of the last resorts. Streptogramin antibiotics are mixtures of two chemically unrelated A and B compounds that act synergistically in vivo against gram-positive pathogens, such as staphylococci, streptococci, and enterococci (Bouilla *et al.*, 1996; Zervos, 1996). Resistance against B compounds is very widespread among enterococci and is mediated

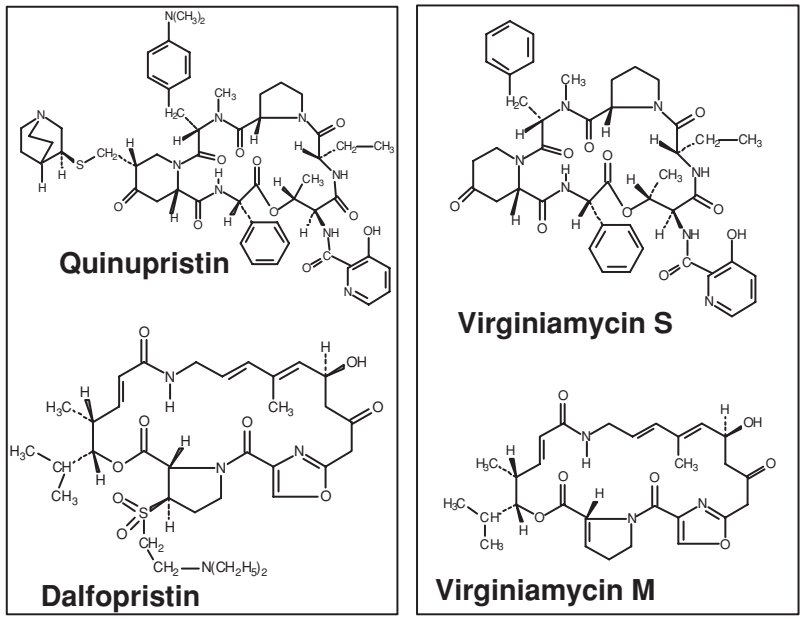


Figure 11.1. Structural relations between quinupristin and virginiamycin S (group B compounds), and dalfopristin and virginiamycin M (group A compounds).

via the *ermB* gene cluster (e.g., on Tn*917*) that confers macrolide-lincosamide-streptogramin B resistance. The synergistic mixture of streptogramins A and B overcomes resistance to B compounds but is inactive in resistance to A compounds. The resistance against streptogramin A compounds in enterococci is mediated by a mechanism involving the streptogramin acetyltransferases SatA and SatG (Werner and Witte, 1999). The streptogramin A compound dalfopristin is structurally related to the streptogramin virginiamycin, which has been used as a growth promoter since 1974 (Figure 11.1). *satA* and *satG* mediates resistance to both virginiamycin and quinupristin/dalfopristin.

Resistance to quinupristin/dalfopristin mediated by *satA* had already been detected in *E. faecium* of clinical origin in Germany before any use in humans. The same resistance gene was found in *E. faecium* from animal feces and from meat products (Werner *et al.*, 1998).

Following this study, we could not demonstrate *satA* in a number of quinupristin/dalfopristin resistant *E. faecium* strains of different ecological origin. In these strains a new enterococcal gene, *satG*, encoding a putative acetyltransferase that confers resistance to streptogramin A compounds and thus to quinupristin/dalfopristin, was identified (Werner et al., 2000).

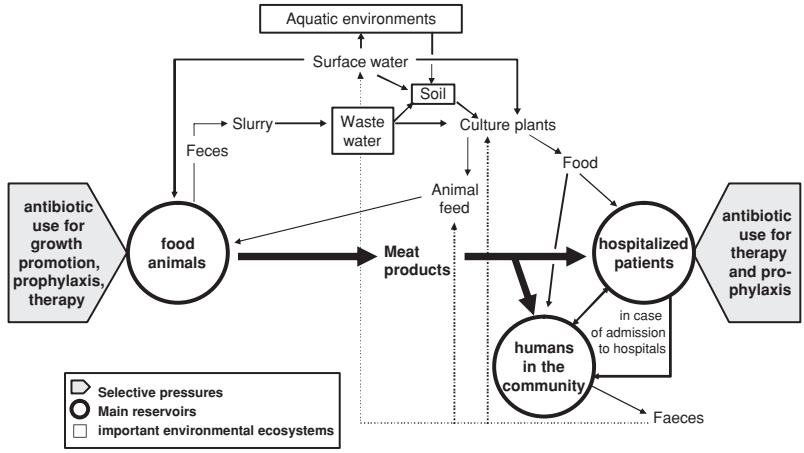


Figure 11.2. Routes of transmission of antibiotic resistant bacteria and resistance genes.

satG was demonstrated in 12 of 46 quinupristin/dalfopristin *E. faecium* of hospital origin (26 possessed *satA*), in six *E. faecium* from 24 poultry carcasses, and in 17 *E. faecium* from 17 samples of poultry manure, thus indicating a reservoir in poultry farms.

There are different possibilities for communication between both large reservoirs of transferable antibiotic resistance, especially in hospitals and in animal husbandry, as shown in Figure 11.2. The food chain is obviously the main route of transmission to humans.

11.2.4. Multi-Resistance Gene Clusters

Because of the continuous use of tylosin in poultry and in pig farming most *E. faecium* isolates are resistant to macrolides. In this species macrolide resistance is exclusively encoded by *erm*B, which is located on Tn*1547* integrated and into conjugative plasmids (Derbise *et al.*, 1997b). In the following examples of multi-resistance, gene clusters containing *ermB* linked to other resistance determinants are described. There is linkage to the *aadE-sat4-aphA-3* cluster, resistance genes conferring resistance to aminoglycosides and to streptothricin (*sat*-4), a group of antibiotics that have been used as growth promoters (see earlier discussion). Besides enterococci, this gene cluster had first been described as a part of transposon Tn*5405* in multi-resistant *Staphylococcus aureus* and coagulase-negative staphylococci (Derbise *et al.*, 1997a; Boerlin *et al.*, 2001), in particular *S. intermedius*. Furthermore, a set of *E. faecium* isolates of animal and human origin was investigated that

had been chosen for deeper analysis because of macrolide and high-level streptomycin resistance phenotypes. In nearly 70% of these isolates a link between the *ermB* and *aadE-sat4-aph3* resistance genes was found. The *aadE-sat4-aphA* cluster was part of Tn*5405* in most of these isolates (Werner *et al.*, 2000). Different types of these gene clusters are shown in Figure 11.3. Type I was the most frequently detected type in isolates from pig, poultry, non-hospitalized humans, and nosocomial infections; this is a further indication of the common resistance gene pool in enterococci of different ecological origin. The cluster of type II contains IS*1216*V, which is mainly known from enterococci. The finding of IS*1216*V in *S. intermedius* containing also *ermB* and *aadE-sat4-aphA* underscores the central role of *E. faecium* as a reservoir for antibiotic resistance genes in Gram-positive species.

11.2.5. The Streptogramin Resistance Gene Cluster

The gene organization on a distinct plasmid fragment was studied in more detail, revealing a genetic linkage of determinants for streptogramin A and B resistance, *vatE* and *ermB*, preceded by an enterococcal-type IS element, a truncated replication gene from small staphylococcal plasmids, and a determinant for chloramphenicol resistance, resulting in a cluster *cat-repB′*-IS*1216V-ermB-vatE*. A screening of 72 *vatE*-type streptogramin-resistant *E. faecium* by polymerase chain reaction (PCR) with overlapping fragments revealed a similar gene arrangement in 16 of these isolates (10 from poultry manure, two from sewage, three from stool samples, and one from a hospitalized patient). At least 45 isolates (59%) possessed a linkage of *vatE* and *ermB*. Ten (13%) of the investigated isolates possessed larger or smaller PCR fragments than described for the reference isolate UW1965 (GenBank accession number AF229200). Sequencing of these amplicons identified smaller fragments due to deletions of the 3′ end of the *repB′* gene and larger fragments due to additional DNA (Figure 11.3, previously unpublished). The *repB* gene encoded an initiator protein for plasmid replication. Whether or not the truncated *repB′* homologues found here are still functional in these isolates has not been investigated. Some streptogramin-resistant *E. faecium* (SREF) possessed additional DNA between IS*1216V* and the *vatE* genes. This included an *ermB* gene cluster identical to the one found in the reference strain UW1965 but flanked by perfect inverted repeats of 62 bp, which could promote a transposition of that cluster. Preceding *vatE*, a new IS element was identified. It was designated IS*1310* and the putative transposase showed 56% identity (72% similarity) at the amino acid level with a transposase from *S. aureus* Mu50. The IS element is flanked by two perfect inverted repeats of

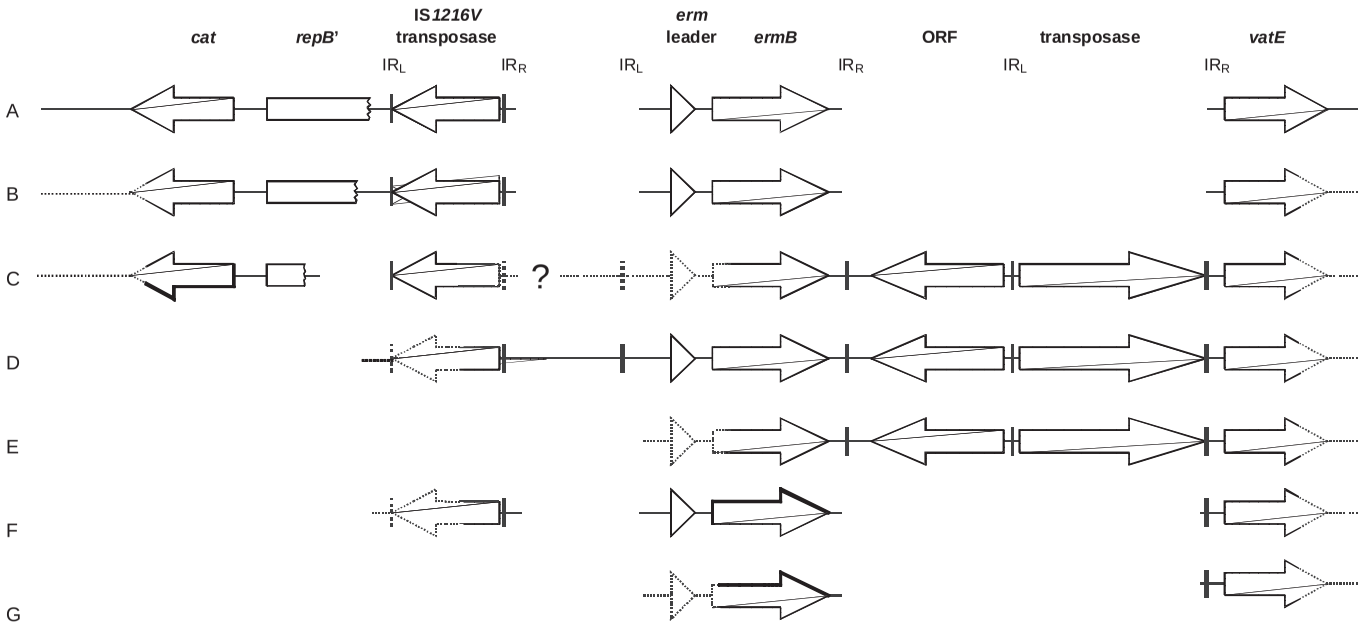


Figure 11.3. Streptogramin resistance gene cluster on different SREF plasmids as deduced from sequencing and polymerase chain reactions. (A) Gene cluster type I (reference isolate UW 1965; 15 additional isolates, see text for details); (B) gene cluster type II (3 of 17 poultry isolates); (C) gene cluster type III (1 of 5 stool sample isolates); (D) gene cluster type IV (2 of 10 sewage and 1 of 1 pork isolates); (E) gene cluster type V (1 of 10 sewage and 2 of 15 stool samples); (f) gene cluster type VI (1 of 17 poultry, 11 of 13 chicken, 4 of 10 sewage, and 4 of 9 patient isolates); (G) gene cluster type VII (2 of 15 stool sample isolates). Dotted lines indicate DNA sequences lying outside the sequenced PCR fragments.

9 bp. Upstream this element an ORF was identified that showed greater than 93% identical amino acids with ORF zeta of *Streptococcus pyogenes* plasmids and ORF 19 of *E. faecalis* plasmid pRE25, respectively (Schwarz *et al.*, 2001). An overview is presented in Figure 11.3.

A multi-resistance gene cluster linking resistance to glycopeptides, macrolides, and aminoglycosides (high-level resistance; Figure 11.4) was detected in a vancomycin-resistant *E. faecium* from a German hospital. Transfer of the plasmid containing this cluster will lead to resistance against major classes of antibiotics used for treating enterococcal infections.

11.2.5.1. Integration of the Glycopeptide Resistance Gene Cluster into Plasmids with a Post-Segregational Killing System

Such kinds of perpetuation of resistance became visible when glycopeptide resistant *E. faecium* strains persisting in Norwegian poultry farms after the ban of avoparcin in 1995 were investigated in more detail. Transposon Tn*1546* was found integrated into the *aadE* gene known from the *S. aureus* plasmid pS194 on non-conjugative plasmids coding for the post-segregational killing system (PSK; ω-ε-ζ, which is known from *Bacillus subtilis*; Ceglowski *et al.*, 1993). The plasmids analyzed contained four intact and four truncated IS-elements that could contribute to subsequent acquisition and dissemination of glycopeptide resistance determinants. That the *vanA*-gene cluster was recently integrated into plasmids with PSK is suggested by an absence of single nucleotide polymorphisms when compared to the originally described Tn*1546* (Sorum *et al.*, 2006). In conclusion, permanent selective pressure resulting from use of avoparcin as a growth promoter had resulted in stable maintenance of glycopeptide resistance in this species.

11.3. CONCLUSIONS

There is strong evidence that antibiotic use in animal feeding had substantially contributed to the creation of a considerable pool of antibiotic resistance genes in *E. faecium*. Their organization in multi-resistance gene clusters with involvement of insertion sequences is the base for co-selection by use of different antibiotic classes. Molecular characterization of resistance genes had revealed that *E. faecium* is a reservoir of such genes. When transferred to different micro-ecological environments by means of food and excretion, resistance can be acquired by other pathogens normally colonizing other ecological niches (e.g., transfer of glycopeptide resistance to *S. aureus*, in particular MRSA; Courvalin, 2006). There is no doubt that introduction of new antibiotics will be followed by resistance development. Besides use

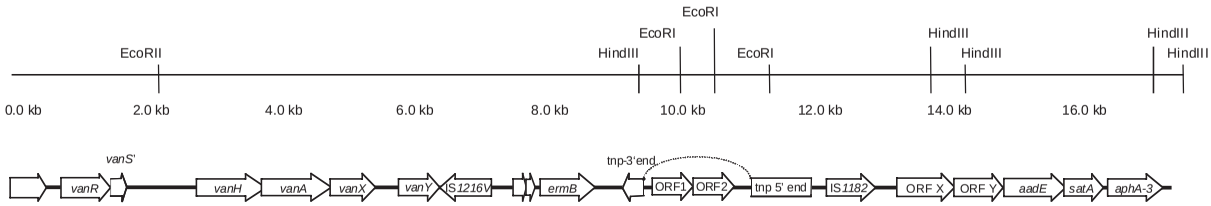


Figure 11.4. A 17.5-kb gene cluster encoding for glycopeptide, macrolide-lincosamide-streptogramin B, and high-level streptomycin, streptothricin, and kanamycin resistance in a hospital *E. faecium* (GenBank accession no. AF51335); putative ORFs resulting from data bank searches.

of an "own potential" expansion or alteration of substrate profiles of own functions, the nearly inexhaustible resistance gene pool in soil microflora (D'Costa *et al.*, 2006) will contribute to this by horizontal gene transfer. As a consequence of human lifestyle *E. faecium* is well equipped for uptake and transfer of additional resistance genes.

As indicated by the discovery of a wide dissemination of class 1 integrons lacking transposition genes as well as antibiotic resistance determinants in bacteria from soil and from sediments (Stokes *et al.*, 2006), antibiotic resistance genes can easily be captured from natural antibiotic producers and spread among bacteria in different microecological systems under selective pressure.

REFERENCES

Aarestrup, F. M. (1995). Occurrence of glycopeptide resistance among *Enterococcus faecium* isolates from conventional and ecological poultry farms. *Microb Drug Resist*, **1**, 255–7.

Arthur, M., and Courvalin, P. (1993). Genetics and mechanisms of glycopeptide resistance in enterococci. *Antimicrob Agents Chemother*, **37**, 1563–71.

Arthur, M., Reynolds, P., and Courvalin, P. (1996). Glycopeptide resistance in enterococci. *Trends Microbiol*, **4**, 401–7.

Boerlin, P., Burnens, A. P., Frey, J., Kuhnert, P., and Nicolet, J. (2001). Molecular epidemiology and genetic linkage of macrolide and aminoglycoside resistance in *Staphylococcus intermedius* of canine origin. *Vet Microbiol*, **79**, 155–69.

Bouilla, H. J., Perri, M. B., Kauffman, C. A., and Zervos, M. J. (1996). Comparative in-vitro activity of quinupristin/dalfopristin against multidrug-resistant *Enterococcus faecium*. *Diagn Microbiol Infect Dis*, **25**, 127–31.

Ceglowski, P., Boitsov, A., Chai, S., and Alonso, J. C. (1993). Analysis of the stabilization system of pSM19035-derived plasmid pBT 233 in *Bacillus subtilis*. *Gene*, **136**, 1–12.

Courvalin, P. (2006). Vancomycin resistance in Gram-positive cocci. *Clin Infect Dis*, **42**(Suppl 1), S25–34.

D'Costa, V. M., McGrann, K. M., Hughes, D. W., and Wright, G. D. (2006). Sampling the antibiotic resistome. *Science*, **311**, 374–7.

Da Silva, A. C., Ferro, J. A., Reinach, F. C., *et al.* (2002). Comparison of the genome of two *Xanthomonas pathogens* with differing host specificities. *Nature*, **417**, 459–63.

Derbise, A., Aubert, S., and El Solh, N. (1997a). Mapping the regions carrying three contiguous antibiotic resistance genes *aadE, sat4,* and *aphA-3* in the genomes of staphylococci. *Antimicrob Agents Chemother,* **41**, 1024–32.

Derbise, A., De Cespedes, G., and El Solh, N. (1997b). Nucleotide sequence of a *Staphylococcus aureus* transposon, Tn*5405,* carrying aminoglycoside resistance genes. *J Basic Microbiol,* **37**, 379–84.

Helmuth, R., and Bulling, E. (Eds.) (1985). *Criteria and methods for the microbiological evaluation of growth promoters in animal feeds.* Berlin: Bundesgesundheitsamt.

Hummel, R., Tschäpe, H., and Witte, W. (1986). Spread of plasmid mediated nourseothricin resistance due to antibiotic use in animal husbandry. *J Basic Microbiol,* **26**, 461–6.

Jansen, T., Forster, P., Levine, M. A., *et al.* (2002). Mitochondrial DNA and the origin of the domestic horse. *Proc Natl Acad Sci USA,* **99**, 10905–10.

Jensen, L. B., Ahrens, P., Dons, L., *et al.* (1998). Molecular analysis of Tn*1546* in *E. faecium* isolated from animals and humans. *J Clin Microbiol,* **36**, 437–42.

Jin, Q., Yuan, Z., Xu, J., *et al.* (2002). Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Res,* **30**, 4432–41.

Klare, I., Heier, H., Claus, H., and Witte, W. (1993). Environmental strains of *Enterococcus faecium* with inducible high level resistance to glycopeptides. *FEMS Microbiol Lett*, **106**, 23–90.

Klare, I., Heier, H., Claus, H., Reissbrodt, R., and Witte, W. (1995a). *vanA*-mediated high-level glycopeptide resistance in *Enterococcus faecium* from animal husbandry. *FEMS Microbiol Lett*, **125**, 165–72.

Klare, I., Heier, H., Claus, H., *et al.* (1995b). *Enterococcus faecium* strains with *vanA*-mediated high-level glycopeptide resistance isolated from animal foodstuffs and fecal samples of humans in the community. *Microb Drug Resist*, **1**, 265–72.

Klare, I., Badstubner, D., Konstabel, C., *et al.* (1999). Decreased incidence of VanA-type vancomycin-resistant enterococci isolated from poultry meat and from fecal samples of humans in the community after discontinuation of avoparcin usage in animal husbandry. *Microb Drug Resist,* **5**, 45–52.

Levy, S. B., Fitzgerald, G. B., and Macone, A. B. (1976a). Spread of antibiotic resistance plasmids from chicken to chicken and from chicken to man. *Nature,* **260**, 40.

Levy, S. B., Fitzgerald, G. B., and Macone, A. B. (1976b). Changes in intestinal flora of farm personnel after introduction of tetracycline-supplemented feed on a farm. *New Engl J Med*, **295**, 583–8.

McKeown, T. (1988). *The origins of human disease.* Oxford: Blackwell.

Miele, A., Bandera, M., and Goldstein, B. P. (1995). Use of primers selective for vancomycin resistance genes to determine *van*-genotype in enterococci and to study gene organisation in *vanA* isolates. *Antimicrob Agents Chemother*, **39**, 1772–8.

Mira, A., Pushker, R., and Rodriguez-Valera, F. (2006). The Neolithic revolution of bacterial genomes. *Trends Microbiol*, **14**, 200–6.

Moran, N. A., and Plague, G. R. (2004). Genomic changes following host restriction in bacteria. *Curr Opin Genet Dev*, **14**, 627–33.

Nierman, W. C., Deshazer, D., Kim, H. S., *et al.* (2004). Structural flexibility in the *Burkholderia mallei* genome. *Proc Natl Acad Sci USA*, **101**, 14246–51.

Parkhill, J., Wren, B. W., Thomson, N. R., *et al.* (2001). Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature*, **413**, 523–7.

Schouten, M. A., Voss, A., and Hoogkamp-Konstantje, J. E. (1997). VRE and meat. *Lancet*, **349**, 1258.

Schwarz, F. V., Perreten, V., and Teuber, M. (2001). Sequence of the 50-kb conjugative multiresistance plasmid pRE25 from *Enterococcus faecalis* RE25. *Plasmid*, **46**, 170–87.

Simonsen, O., Haaheim, H., Dahl, K. H., *et al.* (1998). Transmission of *vanA* type vancomycin-resistant enterococci and *vanA* resistance elements between chicken and humans at avoparcin exposed farms. *Microb Drug Resist*, **4**, 313–8.

Sorum, M., Johnsen, P. J., Aasnes, B., *et al.* (2006). Prevalence, persistence, and molecular characterization of glycopeptide-resistant enterococci in Norwegian poultry and poultry farmers 3 to 8 years after the ban on avoparcin. *Appl Environ Microbiol*, **72**, 516–21.

Stokes, H. W., Nesbo, C. L., Holley, M., *et al.* (2006). Class 1 integrons potentially predating the association with Tn*402*-like transposition genes are present in a sediment microbial community. *J Bacteriol*, **188**, 5722–30.

Tschäpe, H. (1994). The spread of plasmids as a function of bacterial adaptability. *FEMS Microbiol Lett*, **15**, 23–32.

Van Den Braak, N., Van Belkum, A., Keulen, J., *et al.* (1998). Molecular characterization of vancomycin resistant enterococci from hospitalized patients and poultry products in the Netherlands. *J Clin Microbiol*, **36**, 1927–32.

Wagner, A. (2006). Periodic extinction of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomics. *Mol Biol Evol*, **23**, 723–33.

Wei, J., Goldberg, M. B., Burland, V., *et al.* (2003). Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. *Infect Immun*, **71**, 2775–86.

Werner, G., and Witte, W. (1999). Characterization of a new enterococcal gene, *satG*, encoding a putative acetyltransferase conferring resistance to streptogramin A compounds. *Antimicrob Agents Chemother*, **43**, 1813–4.

Werner, G., Klare, I., and Witte, W. (1997). Arrangement of the *vanA* gene cluster in enterococci of different ecological origin. *FEMS Microbiol Lett*, **155**, 55–61.

Werner, G., Klare, I., and Witte, W. (1998). Association between quinupristin/dalfopristin resistance in glycopeptide-resistant *Enterococcus faecium* and the use of additives in animal feed. *Eur J Clin Microbiol Infect Dis*, **17**, 401–2.

Werner, G., Hildebrandt, B., Klare, I., and Witte, W. (2000). Linkage of determinants for streptogramin A, macrolide-lincosamide-streptogramin B, and chloramphenicol resistance on a conjugative plasmid in *Enterococcus faecium* and dissemination of this cluster among streptogramin-resistant enterococci. *Int J Med Microbiol*, **290**, 543–8.

Witte, W., Klare, I., and Werner, G. (2002). Molecular ecological studies on spread of antibiotic resistance genes. *Anim Biotechnol*, **13**, 57–70.

Woodford, N., Adebiyi, A. P., Palepou, M. F., and Cookson, B. (1998). Diversity of *vanA* glycopeptide resistance elements in enterococci from human and nonhuman sources. *Antimicrob Agents Chemother*, **42**, 502–8.

Zervos, M. J. (1996). Vancomycin-resistant *Enterococcus faecium* infections in the ICU and quinupristin/dalfopristin. *New Horizons*, **4**, 385–92.

Part IV Interkingdom Transfer and Endosymbiosis

CHAPTER 12

Eukaryotic Gene Transfer: Adaptation and Replacements

Jan O. Andersson

12.1. INTRODUCTION

The current view of eukaryote phylogeny indicates that microbial eukaryotes (protists) represent the vast majority of the biological diversity within the eukaryotic domain of life, and that multicellular eukaryotes have evolved from protist lineages several times independently (Adl *et al.*, 2005). Accordingly, most eukaryote genome evolution has occurred in microbial lineages, and multicellular eukaryotes with sequestered germlines, such as animals and plants, should be seen as recent evolutionary exceptions, rather than the norm, within the eukaryotic domain (Figure 12.1). Yet, most knowledge about eukaryotic genome evolution comes from multicellular organisms and a few representatives of unicellular lineages, such as yeast. Fortunately, this narrow view of eukaryote genome evolution is now expanding with the advance of genomic sequences from diverse protist lineages.

Several aspects of the lifestyles of protists are more similar to prokaryotic organisms, than to multicellular eukaryotes; differentiation into germ and soma cells are rare, and protists often live in close contacts with cells of distantly related species in the environment. Furthermore, many protists have asexual life cycles, although meiosis appears to be ancestral to eukaryotes (Ramesh *et al.*, 2005). These similarities suggest that protists may share some aspects of genome evolution with prokaryotes. As is obvious from this book, the transfer of genetic material between unrelated lineages, lateral (or horizontal) gene transfer (LGT), is a fundamental mechanism in prokaryotic genome evolution, resulting in genomic plasticity which, for example, provides the basis for adaptation processes (Boucher *et al.*, 2003; Gogarten and Townsend, 2005; Pál *et al.*, 2005). In this chapter the role of LGT in

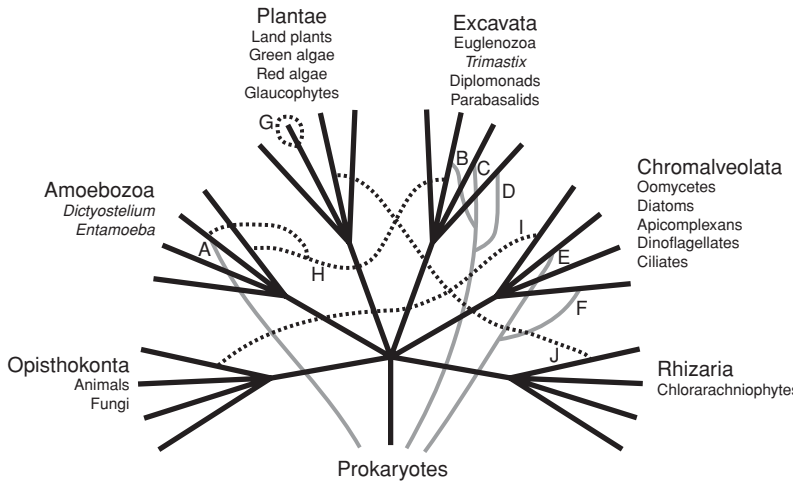


Figure 12.1. Schematic diagram showing the current view of eukaryote phylogeny. The tree is showing the six major eukaryotic supergroups that have been proposed, although the monophyly of some of these remains uncertain (Baldauf, 2003; Simpson and Roger, 2004; Adl *et al.*, 2005). Only names of organismal groups discussed in the text are shown. A–F: Gray lines indicate prokaryote-to-eukaryote lateral gene transfers (LGTs) identified in phylogenomic analyses (Andersson *et al.*, 2007; Carlton *et al.*, 2007; El-Sayed *et al.*, 2005; Ricard *et al.*, 2006; Huang *et al.*, 2004; Loftus *et al.*, 2005). G–J: Dotted lines indicate cases of multiple eukaryote-to-eukaryote gene transfer events (Richards *et al.*, 2006b; Richardson and Palmer, 2006; Archibald *et al.*, 2003; Andersson *et al.*, 2007).

eukaryotes will be discussed with the focus on microbial lineages (Figure 12.1).

12.2. GENE TRANSFER IN EUKARYOTE GENOME EVOLUTION

The accumulation of complete genome sequence data from diverse prokaryotes in the second half of the 1990s rapidly led to the recognition that LGT is an important mechanism in prokaryotic genome evolution (Doolittle, 1999; Ochman *et al.*, 2000). The question as to whether the mechanism also affected microbial eukaryotes could not be answered at the time, because there was a shortage of comparative sequence data from eukaryotes. Nonetheless, there was a tendency to assume that LGT did not play any significant role in eukaryotic genome evolution (Ochman *et al.*, 2000). However, at the turn of the century, reports started to accumulate in the literature in which phylogenetic studies of individual genes showed eukaryotes among prokaryotic sequences, which suggested prokaryote-to-eukaryote

gene transfer events (Doolittle *et al.*, 2003; Richards *et al.*, 2003). More recently, these have been complemented with phylogenetic analyses of genome sequence data. These phylogenomic studies indicate significant prokaryote-to-eukaryote gene transfer in the excavate lineages diplomonads (Andersson *et al.*, 2007), parabasalids (Carlton *et al.*, 2007), and euglenozoans (El-Sayed *et al.*, 2005), the chromalveolate lineage ciliates (Ricard *et al.*, 2006) and apicomplexa (Huang *et al.*, 2004), and the *Entamoeba* lineage (amoebozoa; Loftus *et al.*, 2005; Figure 12.1). Thus, a considerable number of data collectively support the hypothesis that LGT is an important evolutionary mechanism in a number of diverse lineages of microbial eukaryotes. However, putative gene transfer events identified in phylogenomic analyses will inevitably include false positives; many difficulties indeed exist with identifying individual gene transfer events beyond any reasonable doubt (Kurland *et al.*, 2003; Richards *et al.*, 2003; Koonin, 2003; Andersson, 2005). Nevertheless, alternative explanations, such as methodological artefacts and gene duplications and losses, are unlikely to explain all, or even the majority, of the putative cases, as discussed in detail elsewhere (Andersson *et al.*, 2006; Andersson, 2005).

12.3. PROKARYOTE-TO-EUKARYOTE LGT IS AN ONGOING EVOLUTIONARY MECHANISM

To determine the direction and source of a gene transfer event the recipient lineage is required to be nested within the clade representing the donor lineage. In practice this is seldom the case. The reason could be limitations of the phylogenetic methods, or that that the sequence of the closest descendant of the donor lineage is not present in the database, either because it has not been sampled yet, or because it has gone extinct. The problem increases with the number of gene transfers that have occurred within the gene family. Indeed, genes with patchy phyletic distribution (found in only a small proportion of the sampled genomes) often show phylogenetic trees totally at odds with expected organismal phylogenies, frequently with sequences from the three domains of life intermingled (Andersson *et al.*, 2006). This is most likely due to recurring transfers of these genes between distantly related organisms, although the large number of transfer events compared to the relatively limited number of taxa makes detailed interpretations of donor lineages problematic. As a consequence, only a donor phylum or even domain may be possible to identify in many cases in inter-domain LGT events.

For the genes where a direction of the transfer between prokaryotes and eukaryotes has been possible to identify, the vast majority are in the

direction to eukaryotes. A transfer of tubulin genes to *Prosthecobacter* and a few putative transfers of metabolic genes are rare exceptions (Jenkins *et al.*, 2002; Andersson *et al.*, 2003; Schlieper *et al.*, 2005; Richards *et al.*, 2006a). The shortage of eukaryote-to-bacteria transfer is puzzling. The large number of prokaryotes compared to eukaryotes in most environments could make such transfers rather unlikely compared to a transfer in the other direction, and the presence of introns in many eukaryotic genes could also represent a barrier for transfers to prokaryotes (Andersson, 2005). At any rate, the data indicate that there is a flow of genes from prokaryotic organisms to a number of eukaryotic lineages (Figure 12.1).

In a phylogenomic analysis of the complete genome sequence of the parabasalid *Trichomonas vaginalis* (a human parasite), 152 genes were identified as possible cases of gene transfers from prokaryotic organisms. Interestingly, many of the genes appeared to have been transferred from the lineages within the Bacteroidetes phylum of bacteria, which are frequent inhabitants of the vertebrate intestinal flora and therefore likely to come in contact with parasitic protists (Carlton *et al.*, 2007). This illustrates a general pattern in prokaryote-to-eukaryote gene transfer; in cases where a donor lineage is possible to identify, a transfer in an environment similar to the ones of the extant descendants of the donor and recipient organisms is generally indicated. A strong bias towards donors belonging to the Bacteroidetes phylum was indeed also observed among the 96 putative prokaryote-to-eukaryote gene transfers identified in the analyses of the human intestinal parasite *Entamoeba histolytica* genome sequence (Ricard *et al.*, 2006). Similarly, the phylogenetic analyses of expressed sequenced tags (ESTs). from rumen ciliates identified members of Firmicutes as the donor lineage in bacteria-to-ciliate gene transfers in 34 of the 133 cases; Firmicutes have been shown to dominate in the rumen bacterial ecosystem (Edwards *et al.*, 2004).

A pattern of gene transfer between ancestors of present-day organisms found in similar environments should not be surprising since physical proximity of the donors and recipients is expected to greatly increase the probability that successful gene transfers will occur. This makes biological sense and therefore argues against the possibility that putative cases of gene transfer inferred from genomic sequences are all due to artefacts in the phylogenetic methods or to differential gene loss scenarios; these factors have indeed previously led to false indications of LGT in eukaryotes (Genereux and Logsdon, 2003; Richards *et al.*, 2003). The pattern also suggests that gene acquisitions are ongoing in the evolution of eukaryotes; the genes were likely acquired after the eukaryotes diverged and adapted to different environments. In agreement with this view, homologues to the LGT genes identified in rumen

ciliates were typically absent from the completely sequenced free-living ciliate *Tetrahymena thermophila*, suggesting that addition of these genes to some ciliate lineages is more recent than the diversification of ciliates (Ricard *et al.*, 2006).

12.4. ENDOSYMBIOTIC GENE TRANSFER

Endosymbiotic gene transfer (EGT) is a special case of LGT that occurs between an engulfed cell that has evolved into an endosymbiont and the host nucleus. The source of the gene is an organelle with an endosymbiotic origin (i.e., the mitochondrion or the plastid) rather than an organism. EGT may have resulted in the influx of a large number of bacterial genes into eukaryotic genomes; it has indeed been proposed that the majority of yeast genes could have originated from the endosymbiont that gave rise to the mitochondrion (Esser *et al.*, 2004). Similarly, 18% of the genes in the plant *Arabidopsis thaliana* have been suggested to have a cyanobacterial origin via the plastid (Martin *et al.*, 2002), and a recent analysis indicated 132 genes of cyanobacterial origin in the genome of the glaucophyte alga *Cyanophora paradoxa*, corresponding to 10.8% of the analyzed genes (Reyes-Prieto *et al.*, 2006). These results indicate that the genetic contribution from the endosymbionts might go far beyond the functions performed by the organelles. However, a cyanobacterial origin for a nucleus-encoded gene in a plastid-containing eukaryote does not necessarily mean that it has an endosymbiotic origin; cyanobacteria and algae share niches and may therefore exchange genes. Indeed, in a phylogenomic analysis on the anaerobic protist *Spironucleus salmonicida* (a diplomonad) that was similar to the one performed on the glaucophyte alga, 11.8% of the genes were found to have originated from prokaryotic sources after the divergence of eukaryotes (Andersson *et al.*, 2007); these transfers most likely did not involve any organelles with endosymbiotic origin. Thus, examples of a large fraction of genes with prokaryotic origin are not found only in plastid-containing eukaryotes, suggesting that some of the genes with cyanobacterial origin (Martin *et al.*, 2002; Reyes-Prieto *et al.*, 2006) could be the result of LGT rather than EGT. There are indeed examples of gene acquisitions from cyanobacteria of genes encoding functions performed in the plastids of algae (Waller *et al.*, 2006); such results probably would have been interpreted as an EGT in a genome-wide analysis. Obviously, the drastic contribution to the eukaryotic nucleus from the organelles with bacterial endosymbiotic origin that has been proposed (Martin *et al.*, 2002; Reyes-Prieto *et al.*, 2006; Esser *et al.*, 2004) needs to be further evaluated taking the possibility of LGT into account.

Some algal lineages have plastids of secondary endosymbiotic origins; they have engulfed a eukaryotic alga which has been enslaved and incorporated as a plastid (Archibald and Keeling, 2002). EGT between the nucleus of the enslaved photosynthetic eukaryotic to the host nucleus will result in eukaryote-to-eukaryote gene replacements (Archibald and Keeling, 2002). Chlorarachniophytes are an example of an algal group that has acquired plastids secondarily. The nucleus of the symbiont is maintained as a highly reduced nucleomorph. Therefore, it is expected that many genes have been transferred from the nucleomorph to the host nucleus, including genes with cyanobacterial origin that are coding for proteins used in the plastid. This indeed turned out to be the case; the majority of the studied genes encoding proteins targeted to the plastid in *Bigelowiella natans* (a chlorarachniophyte) had a green algal origin, as expected (Archibald and Keeling, 2002). However, the phylogenies for 21% of the studied genes showed topologies that strongly indicated an origin via gene transfer from sources distinct from the secondarily derived plastid. Most of these LGTs were of red algal origin (Figure 12.1), and a few were even of bacterial origin (Archibald and Keeling, 2002). This indicates that functional replacements of genes in the symbiont nucleus by genes from other sources occur at an appreciable frequency, as a complement to EGT.

12.5. LATERAL GENE TRANSFER LEADS TO ADAPTATION

LGT has been identified to be a major evolutionary mechanism in niche adaptation in prokaryotic organisms (Boucher *et al.*, 2003; Pál *et al.*, 2005; Gogarten and Townsend, 2005; Beiko *et al.*, 2005). Closely related isolates often differ substantially in gene content, which could be explained by niche adaptation. For example, the majority of the genes found in either of two *Prochlorococcus* isolates optimized to a life at different depth in the ocean were unique to one or the other of the isolates (Rocap *et al.*, 2003). The contents of genomes of prokaryotic organisms (with the possible exception of symbionts and parasites) are thought to be constantly changing throughout evolutionary time, with only a minority of genes that is very rarely transferred (a core) and a larger fraction of genes that is more frequently exchanged (a shell). The core genes are mainly encoding proteins in the basic cellular machinery, while the shell genes are providing the organism with functions resulting in niche adaptation (Pál *et al.*, 2005; Gogarten and Townsend, 2005; Beiko *et al.*, 2005). Thus, there probably exists a number of gene families that are only narrowly distributed via frequent gene transfers among organisms sharing a specific niche. Analyses of sequences from microbial

eukaryotes have suggested that this gene sharing might not be limited to prokaryotic organisms.

An early example of niche adaptation in eukaryotes by gene acquisition from bacteria came from rumen fungi; their cellulose-degrading enzymes show similarities to bacterial enzymes, suggesting that they were introduced by gene transfer (Garcia-Vallve *et al.*, 2000). These uptakes likely were key events in the colonization of the rumen by fungi (Garcia-Vallve *et al.*, 2000). Interestingly, adaptation to an anaerobic carbohydrate-rich environment by gene acquisitions in a larger scale was invoked to explain the observation that rumen ciliates shared a substantial number of genes with anaerobic prokaryotes, many of which function in carbohydrate metabolism (Ricard *et al.*, 2006). This suggests that the rumen was first colonized by prokaryotes and secondarily independently by distantly related microbial eukaryotes, likely facilitated by prokaryote-to-eukaryote LGT.

Giardia lamblia (a diplomonad) and *Entamoeba histolytica* are two human intestinal parasites belonging to different eukaryotic groups: excavata and amoebozoa, respectively (Adl *et al.*, 2005; Figure 12.1). They have adapted to a life as parasites in anoxic environments independently; their common ancestor was likely an aerobic free-living protist (Cavalier-Smith, 2002). Phylogenetic analyses of genes have indeed suggested that this adaptation at least partly is due to acquisition of genes from anaerobic prokaryotes; the same genes were transferred independently to the two lineages (Andersson *et al.*, 2003). Increased taxon sampling of a few of these genes resulted in the interpretation of a larger number of gene transfer events to a diversity of protist lineages living in anoxic environments (Andersson *et al.*, 2006). Thus, these are examples of niche-specific genes with patchy phyletic distribution that are exchanged between all three domains of life leading to adaptation. Genome-wide analyses of *E. histolytica* and the fish pathogen *Spironucleus salmonicida* (another diplomonad) have shown that this is not restricted to a few genes. Rather, a large number of mostly metabolic genes have been identified to have originated via LGT, likely providing metabolic adaptation to the niches of these protists (Loftus *et al.*, 2005; Andersson *et al.*, 2007, 2006).

A more direct approach to study the adaptation process in eukaryotes was taken to determine whether fungi have acquired genes from their environment (Wenzl *et al.*, 2005). Glucuronides are present in vertebrate urine and it has been known that some soil bacteria are able to utilize this compound as a carbon source. Wenzl *et al.* hypothesized that fungi would also benefit from using glucuronides and set up a functional screen where they successfully isolated two ascomycetes from soil capable of degrading the compound.

They also obtained several bacterial isolates using the same approach. Phylogenetic analyses of the β-glucuronidase genes suggested that these had been transferred from an ancestor of the isolated bacteria to a fungal lineage. The use of a functional screen provided evidence for an expanded metabolic capacity of the recipient after the uptake of the gene, which is beneficial in soil if vertebrate urine is present (Wenzl *et al.*, 2005). Another example of experimental evidence for niche adaptation via gene acquisition in fungi comes from yeast (Gojkovic *et al.*, 2004; Hall *et al.*, 2005). It was shown that a functional gene encoding the enzyme dihydroorotate dehydrogenase derived from an ancestor of *Lactococcus* was required for anaerobic synthesis of uracil in *S. cerevisiae* (a facultative anaerobe). The ancestral fungal dihydroorotate dehydrogenase could complement the derived bacterial gene under aerobic but not anaerobic conditions. Furthermore, experimental transfer of the bacteria-type dihydroorotate dehydrogenase from *S. cerevisiae* to *Pichia stipitis* transformed the latter from an aerobe to a facultative anaerobe (Shi and Jeffries, 1998). Together these observations strongly suggest that the incorporation of the bacterial gene in the yeast genome resulted in a facultative anaerobic lifestyle (Gojkovic *et al.*, 2004; Hall *et al.*, 2005).

Oomycetes is a group of eukaryotic plant pathogens that resembles osmotrophic filamentous fungi, but belongs to a group without any phylogenetic affinity with fungi (Adl *et al.*, 2005; Figure 12.1). Thus, the similarities in lifestyle between oomycetes and parasitic fungi are due to convergent evolution rather than shared ancestry (Latijnhouwers *et al.*, 2003). Consequently, genes shared between the groups to the exclusion of other eukaryotes, which could explain the similarities, are likely to represent adaptive gene exchanges. Indeed, bioinformatic analyses indicated a dozen genes that showed this pattern, four of which were shown to have been transferred from fungi to oomycetes (Richards *et al.*, 2006b; Figure 12.1). These genes encoded function related to the utilization of rare metabolites that could be coupled to an osmotrophic lifestyle (Richards *et al.*, 2006b; Andersson, 2006a). Interestingly, three of the four genes studied with phylogenetic methods indicated bacteria-to-fungi gene transfers preceding the fungi-to-oomycete transfers, suggesting a gene flow from soil bacteria to fungi and then to oomycetes.

12.6. EVOLUTION OF PATHOGENICITY IN EUKARYOTES BY GENE ACQUISITION

The evolution of pathogenicity could be seen as an adaptation of the parasite to explore the host organism. In bacteria, evolution of pathogenicity

is often coupled with changes in the gene repertoire, and reports have started to appear indicating a similar pattern for eukaryotic pathogens (Lawrence, 2005). For example, the filamentous fungus *Nectria haematococca* has a cluster of genes that is dispensable for growth in culture but required for the fungus to cause disease on pea (Temporini and VanEtten, 2004). The gene cluster is absent from closely related fungi but present in the distantly related pea pathogen *Fusarium oxysporum*. The patchy distribution of the gene cluster and the fact that it is required for pathogenicity on pea suggests that it has been distributed via gene transfer between these pathogens; the uptake of the cluster by *N. haematococca* likely conferred the ability to colonize peas and cause disease (Temporini and VanEtten, 2004).

The emergence of a disease on wheat in the first half of the last century is explained by an even more striking example: a gene transfer between two fungi (Friesen *et al.*, 2006). A highly virulent form of *Pyrenophora tritici-repentis* that causes tan spot disease, a serious problem on wheat, has been rapidly spreading around the world after its first occurrence in the United States in 1941. This fungus species had been known to be pathogenic on wheat, but it previously caused only a mild disease. A single gene, *toxA*, encoding a host-specific toxin (ToxA) has been identified to be responsible for the virulence on wheat in the aggressive strain (Ciuffetti *et al.*, 1997). Intriguingly, a gene with 99.7% identity to the *P. tritici-repentis toxA* gene was identified in the genomic sequence of *Stagonospora noderum*, another pathogen on wheat (Friesen *et al.*, 2006). This is highly unexpected since these two pathogens are not closely related, and the expected level of sequence identity between genes in their genomes is around 80%. Furthermore, no *toxA* genes could be found in any of the more closely related species (pathogens to other crops), consistent with a recent gene transfer between the wheat pathogens. Comparison of the *toxA* loci in the two species identified an 11-kb region of highly similar sequence including a transposase gene (Friesen *et al.*, 2006). Two observations supported a direction of the transfer from the *S. noderum* to *P. tritici-repentis*. First, *S. noderum* has been known as a serious wheat pathogen for a much longer time than *P. tritici-repentis*. Second, only a single *toxA* haplotype was found among the 54 *P. tritici-repentis* isolated from diverse geographical populations, while 11 haplotypes was found among the 95 sampled *S. noderum* pathogen populations. Although this likely represents a gene transfer affecting a fungus in historical time, the mechanism remains elusive (Friesen *et al.*, 2006; Sanders, 2006).

To be able to penetrate unwounded host tissue, plant pathogens have to be able to degrade cutin, which is an insoluble lipid polyester that forms a major part of the plant cuticle. Accordingly, functional cutinase enzymes have

been shown to be critical for pathogenicity on plants in fungi (Kolattukudy *et al.*, 1995). Interestingly, bioinformatic analyses have identified cutinase homologues in only two additional organismal groups: actinobacteria and oomycetes. Functional analyses of a cutinase from *Phytophthora brassicae* (an oomycete) showed it to be expressed early during infection of the host *Arabidopsis thaliana* (Belbahri *et al.*, 2008). Phylogenetic analyses identified an actinobacterial origin for the oomycete cutinases, indicating that the virulence of this group of important plant pathogens at least to some extent is the result of an inter-domain gene transfer event. Given that bacteria-to-eukaryote gene transfers appear more common than transfers in the reverse direction, it seems plausible that the fungal virulence factor originated via a gene acquisition from bacteria, although the sequence analyses are inconclusive as to the group in which the cutinase gene originated (Belbahri *et al.*, 2008). At any rate, the origin of the cutinase enzyme is almost certainly more recent than the diversification of the three organismal groups where it has been detected; cutin is unique to land plants and probably evolved as an adaptation to life on land (Kenrick and Crane, 1997).

As mentioned earlier, anaerobic protists have adapted to a life in anoxic environments at least partly via gene acquisitions from prokaryotes (Andersson *et al.*, 2003, 2006, 2007; Loftus *et al.*, 2005; Andersson, 2006b). The transformation to an anaerobic metabolism probably enabled them to explore the intestine of vertebrates. For example, both commensals and pathogens of fish are found within the genus *Spironucleus,* diplomonads that are adapted to a life at low oxygen pressure (Jørgensen and Sterud, 2007). Although the genetic basis for their different lifestyles within the host is unknown, genes of prokaryotic origin seem to play a role in the virulence of diplomonad pathogens. Rubrerythrins, A-type flavoproteins, and arginine deiminase are proteins that protect other microbes against reactive oxygen species and nitric oxide, which are important factors in the host's protection against mucosal pathogens (Roxström-Lindquist *et al.*, 2006; Sarti *et al.*, 2004; Sztukowska *et al.*, 2002). The diplomonad homologues of these genes have apparently been acquired from prokaryotes via LGT (Andersson *et al.*, 2007). However, also free-living relatives of the parasites encode at least some of these enzymes (Biagini *et al.*, 2003), suggesting that other properties than their host protection roles likely were the selective basis for the fixation of these genes. For example, the introduction of the three genes coding for the arginine dihydrolase pathway, including the gene for arginine deiminase (Andersson *et al.*, 2007), enabled the diplomonads to use amino acids as an energy source (Adam, 2001), especially under limited oxygen conditions (Biagini *et al.*, 2003). Nevertheless, it cannot be excluded that some virulence

factors of these protists have been acquired by gene transfer from other organisms.

12.7. FUNCTIONAL REPLACEMENTS

The LGT is not limited to genes that confer novel function on the recipient lineage. Functional replacements of genes do also occur where there is no obvious indication of a selective advantage. A classic example of this is the genes coding for the 20 tRNA synthetases. The encoded proteins have to be present in the cell to have a functional protein synthesis, and yet the genes have been shown to undergo relatively frequent gene transfers (Wolf *et al.*, 1999). Such functional replacements of tRNA-synthetase genes have been shown not to be restricted to prokaryotes. For example, an ancestor of diplomonads and parabasalids has acquired archaeal prolyl- and alanyl-tRNA synthetase genes (Andersson *et al.*, 2005), and an ancestral opisthokont replaced the eukaryotic tyrosyl-tRNA synthetase with a haloarchaeal homologue (Huang *et al.*, 2005).

Phylogenetic studies of genes coding for metabolic functions that are present in the majority of organisms have suggested that functional replacements indeed might be occurring at frequencies in microbial eukaryotes that are similar to the frequencies in prokaryotic organisms for these genes. For example, four distinct protein families of glutamate dehydrogenase (GDH) have been identified to date, and the distribution of these in the three domains of life superficially indicates that all four versions were present in the last common universal ancestor (Andersson and Roger, 2003). However, no extant genome contains genes coding for members of more than two of the families. Therefore, it seems rather unlikely that ancestral genomes maintained genes for all four GDH families for long evolutionary times, as required to explain the current distribution of the genes in the absence of gene replacements. Phylogenetic analyses of the four families indeed identified numerous gene transfers, affecting both prokaryotic and eukaryotic organisms. Interestingly, there were no indications that gene replacements are more frequent in prokaryotes than in microbial eukaryotes within these gene families (Andersson and Roger, 2003).

Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) is a key enzyme in glycolysis as well as in the Calvin cycle. The gene coding for GAPDH was most likely present in the last common eukaryotic ancestor, and it appears to be a reasonable assumption that it has been maintained throughout eukaryotic evolution via vertical inheritance. If so, it would be a suitable phylogenetic marker for organismal phylogeny. However, it was recognized early

on that phylogenetic trees based on GAPDH indicated a complex picture of the evolution of the gene with several distinct prokaryotic origins of the gene in different eukaryotic lineages (Martin *et al.*, 1993). As more GAPDH sequences were accumulating in the databases, the phylogenies became even more confusing with eukaryotes clearly polyphyletic. For example, euglenozoans and parabasalids were grouping with bacterial lineages, such as spirochetes. These results strongly suggested inter-domain gene replacements via LGT in eukaryotes (Figge and Cerff, 2001). Strikingly, additional sampling of euglenozoan GAPDH identified two more bacterial-type genes within this eukaryotic group gene (Qian and Keeling, 2001). Gene replacements within this gene family apparently are not restricted to inter-domain LGTs; analyses of dinoflagellate sequences suggested transfers between algal lineages (Takishita *et al.*, 2003), and excavates appear to have exchanged GAPDH genes (Stechmann *et al.*, 2006), further illustrating the complex evolutionary history of this gene family.

An analysis of the evolutionary origins of the genes encoding glycolytic enzymes in the anaerobic flagellate *Trimastix pyriformis* provided intricate results, showing that a complex evolutionary history is not restricted to GAPDH within the glycolytic pathway (Stechmann *et al.*, 2006). While a few genes indicated monophyletic eukaryotes without any strong indications of gene exchanges, others showed phylogenies totally at odds with accepted organismal relationships, which only could be explained by multiple gene transfers affecting protists. Accordingly, analyses of four of the *Trimastix* genes strongly suggested origins via LGT, while two showed weak indications of transfers, and four were probably vertically inherited. The sources of the transfers could not be determined, probably because of frequent transfers of the genes in general, although the data pointed at bacterial origins for most of the genes (Stechmann *et al.*, 2006). Nevertheless, unexpected sister relationships between eukaryotic lineages in a few cases suggested prokaryote-to-protist transfers followed by intra-domain gene transfers. Still, the most striking result from this rigorous analysis of a single enzymatic pathway was the huge differences in the frequencies of LGT between the individual enzymes. It was speculated that there indeed was selection for replacements of the frequently transferred genes, although the basis for such an selection is elusive in the absence of more detailed knowledge of the enzymatic properties of various versions of the enzymes (Stechmann *et al.*, 2006).

In a similar study, Richards and colleagues set out to determine the evolutionary origin of the shikimate pathway, which produces chorismate, a precursor of many aromatic compounds, from erythrose 4-phosphate and

phosphoenol pyruvate in seven steps (Richards *et al.*, 2006a). They found that the pathway is present in at least three of the six eukaryotic supergroups: opisthokonts (fungi, but not metazoa), plantae, and chromalveolates. Interestingly, the eukaryotic sequences did not form a monophyletic group in any of the analyzed individual enzymes; the plantae sequences branch with eubacterial sequences separate from the fungi and most of the chromoalveolate sequences. At face value, a plastid origin of the plantae shikimate pathway could explain such a pattern – the plant enzymes are mainly active in the chloroplast. However, the details of the phylogenetic data argue against such a scenario; only two of the seven plant shikimate pathway enzymes branches with cyanobacterial homologues (Richards *et al.*, 2006a). The other enzymes branch with different eubacterial lineages, suggesting either a stepwise acquisition of the pathway in plantae from diverse bacterial sources, or gene replacements via eubacterial-to-plantae LGTs after the introduction of the pathway from the plastid endosymbiont via EGT. The origins of the pathway in the other two eukaryotic supergroups are also difficult to determine from the data. The fungal and chromalveolates sequences branches together in the analyses where they are both present (Richards *et al.*, 2006a). This could be explained by either the presence of the pathway in the common ancestor of eukaryotes followed by loss in many lineages, or transfer of the individual enzymes between fungi and chromalveolates. Both of these scenarios are problematic. The authors favour the former and provide a detailed model about the evolution of the pathway in eukaryotes (Richards *et al.*, 2006a). However, five of the genes in the pathway are fused in some of these lineages, suggesting that a limited number of transfer events could explain the monophyly of fungi and chromalveolates; it was indeed recently shown that oomycetes (chromalveolates) have acquired several genes from fungi (Richards *et al.*, 2006b).

12.8. WHAT ABOUT TRANSFERS BETWEEN EUKARYOTES?

Some of the mentioned gene transfers seem to have involved gene exchanges between eukaryotic lineages which do not involve an endosymbiotic organelle (Table 12.1). For example, fungi exchanged a virulence factor (Friesen *et al.*, 2006), oomycetes acquired genes from fungi (Richards *et al.*, 2006b), and genes were transferred between algal lineages (Archibald *et al.*, 2003; Figure 12.1). Still, the majority of the identified gene transfer events affecting microbial eukaryotes do seem to represent inter-domain transfers, despite the fact that intra-domain transfers appear more straightforward mechanistically. The reason for this is unclear; it could simply be

Table 12.1. *Putative intra-domain gene transfers affecting eukaryotes*

Gene/protein	Gene exchange	Reference
ToxA	Fungi-to-fungi	Friesen *et al.*, 2006
Four genes	Fungi-to-oomycete	Richards *et al.*, 2006b
flp gene	Between animals (*Hydra*) and parabasalids	Steele *et al.*, 2004
Enolase	Diatom-to-dinoflagellate	Harper and Keeling, 2004
GAPDH	Euglenozoa-to-dinoflagellate	Takishita *et al.*, 2003
GAPDH	Between *Trimastix* and parabasalid	Stechmann *et al.*, 2006
Alanyl-tRNA synthetase	Excavate-to-ciliate and excavate-to-*Entamoeba*	Andersson *et al.*, 2005
Glucosamine-6-phosphate isomerase	Amoebozoa-to-ciliate	Andersson *et al.*, 2006
12 genes	Excavate-to-amoebozoa	Andersson *et al.*, 2007
Threonine dehydratase	Between diplomonads and *Dictyostelium*	Andersson *et al.*, 2003
Several plastid-targeted proteins	Various algal lineages to Chlorarachniophytes	Archibald *et al.*, 2003
EFL/EF1a	Gene replacements in several lineages	Keeling and Inagaki, 2004; Gile *et al.*, 2006
Various mitochondrial genes	Gene replacements between unrelated plants	Richardson and Palmer, 2006

that eukaryote-to-eukaryote transfers are more difficult to detect because of the relative lack of genome sequence data from eukaryotes compared to prokaryotes. If so, the reported intra-domain transfers (Table 12.1) could represent the tip of an iceberg, rather than a few exceptions to the rule. An unexpected indication that this might be the case has come from the analyses of plant mitochondria. It was found that some mitochondrial genes in

flowering plants produced phylogenies totally at odds with organismal relationships, which were inferred as frequent gene transfer events between distantly related plant species (Bergthorsson *et al.*, 2003). Subsequent analyses have shown this phenomenon to be widespread within plant mitochondria (Richardson and Palmer, 2006; Figure 12.1), but very rare in plant plastids (Rice and Palmer, 2006).

There are accumulating data suggesting that microbial eukaryotes also exchange genes at an appreciable rate (Table 12.1). A striking example comes from the studies of the phylogeny and distribution of elongation factor EF-1α (Keeling and Inagaki, 2004). Surprisingly, some lineages apparently lack this protein, which is essential for protein synthesis. However, in these lineages a functionally similar, but evolutionary distinct, EF-like protein (EFL) is present which most likely has replaced the canonical EF-1α. The complex pattern of presence or absence of EF-1α and EFL – they are rarely found in the same organisms – suggests that EFL has replaced EF-1α several times independently in eukaryotes (Keeling and Inagaki, 2004). Detailed analyses of the evolution of the two genes in one of the major eukaryotic groups, the chromalveolates, indeed suggest three independent replacements of EF-1α by EFL (Gile *et al.*, 2006). At any rate, the gene encoding EFL appears to be extremely mobile in eukaryotic evolution. A selective advantage to use EFL rather than EF-1α would explain this mobility and suggests that EFL, with a relatively recent evolutionary origin, therefore slowly is replacing EF-1α in eukaryotes over evolutionary time (Keeling and Inagaki, 2004). However, the functional basis for such a selection, if it exists, is currently unknown.

Many of the reported cases of eukaryote-to-eukaryote transfer have been detected in analyses targeted at inter-domain gene transfers (Table 12.1). For example, alanyl-tRNA synthetase was identified to be transferred from Archaea to an ancestor of diplomonads and parabasalids, and the gene apparently was further distributed to the lineages leading to ciliates and *Entamoeba* in two intra-domain transfers after the initial inter-domain transfer (Andersson *et al.*, 2005). Furthermore, in a phylogenomic approach to detect transfers affecting diplomonad genes, intra-domain transfers that happened after an initial inter-domain transfer were detected in 13 of 68 cases. Of these, 12 were gene exchanges with the amoebozoan lineages leading to *Entamoeba* and *Dictyostelium*, respectively (Andersson *et al.*, 2007; Figure 12.1). Given that the probability for a functional gene expression followed by translation into a protein should be higher for a gene acquired from a eukaryote than for one acquired from a prokaryote, it could be that genes acquired from prokaryotes and, once adapted to a eukaryotic, genome tend to subsequently

spread among eukaryotes via intra-domain LGTs (Andersson *et al.*, 2007). If so, groupings of unrelated eukaryotes are expected to be observed nested within prokaryotes in phylogenetic trees of genes that have been affected by both inter- and intra-domain transfers. A consequence would be that apparently shared genes derived from prokaryote-to-eukaryote gene transfer events should be treated with caution as phylogenetic markers (Andersson *et al.*, 2005; Huang *et al.*, 2005; Huang and Gogarten, 2006) since they could represent intra-domain transfers.

12.9. PATTERNS OF GENE TRANSFERS IN EUKARYOTES MAY INDICATE MECHANISMS

Common to all cases of gene transfer affecting eukaryotes is that a clear picture of the mechanism that enable the uptake of foreign genes is lacking, although the quantities of data currently available from a diversity of lineages strongly suggest that such mechanisms do exist. A variety of speculative mechanisms have indeed been suggested, such as phagocytosis, symbiosis, introgression via interspecies hybrids, and transfection via viruses (Gogarten, 2003). Perhaps the most substantial among these is the suggestion that replacements of ancestral genes over evolutionary time with genes from food organisms is an unavoidable consequence of a phagotrophic lifestyle in eukaryotic lineages (Doolittle, 1998). LGT indeed appears to be common in phagotrophs (Andersson, 2005; Archibald *et al.*, 2003), although relatively high frequencies have also been observed in non-phagotrophic lineages (Andersson, 2005), suggesting either that gene acquisitions occurred in phagotrophic ancestors or, probably more likely, that there are other significant mechanisms for gene transfer in eukaryotes. For example, the recent transfer of virulence factors between fungi (Friesen *et al.*, 2006), which have been osmotrophic for a long evolutionary time, could indicate that incompatibility barriers are more leaky than had been thought, which could lead to gene transfers via anastomosis of mycelia from unrelated fungi (Sanders, 2006). Similar mechanisms, or transfection via viruses, were suggested to explain the transfers between fungi and oomycetes (Richards *et al.*, 2006b). Strikingly, abundant bacterial-type genes have been identified in some eukaryotic viruses that infect eukaryotes, which suggests that they indeed could be vectors for inter-domain gene transfers (Filee *et al.*, 2007).

Gene transfers are events that happened in evolutionary times. Although studies of mechanisms by which genes can be successfully transferred between modern organisms are of great value, the evolutionary significance of these mechanisms is hard to assess. Therefore, such studies need to be

complemented with comparative studies of gene transfer. One such study of different lineages of amoebozoa identified variation in the frequencies of eubacterium-to-eukaryote LGTs (Watkins and Gray, 2006). If such results are correlated with variations of other properties, such as feeding modes, symbiotic relationships, and range of infecting viruses, the understanding of gene transfer mechanisms in eukaryotes might possibly be increased.

12.10. CONCLUDING REMARKS

The number of reports of gene transfers affecting eukaryotes is steadily increasing, indicating that the field has passed the stage when the main question was whether gene transfer is an extremely rare exception or a significant mechanism in the evolution of eukaryotes. However, the data indicate that there are huge variations in the role of LGT in different lineages, in relation to EGT, gene duplication, and gene loss. Similarly, large variations in the frequencies of gene transfers affecting eukaryotes have been observed between genes, even if they function in the same metabolic pathway. A more detailed knowledge of the variation of the biochemistries of the homologous enzymes and their roles within the cells is obviously needed for an understanding of these variations (Stechmann *et al.*, 2006). Finally, most studies of LGT in eukaryotes have focused on prokaryote-to-eukaryote gene transfer, most likely because it is easier to screen for these than for eukaryote-to-eukaryote transfers; intra-domain transfers are indeed detected if they are taken into account (Andersson *et al.*, 2007). This suggests that future studies might show that eukaryote-to-eukaryote (Table 12.1) as well as prokaryote-to-eukaryote gene transfers are frequent in eukaryote genome evolution.

ACKNOWLEDGMENTS

I thank Lassaad Belbahri (University of Applied Sciences of Western Switzerland) for sharing results prior to publication. The author is supported by a Swedish Research Council (VR) Grant.

REFERENCES

Adam, R. D. (2001). Biology of *Giardia lamblia*. *Clin Microbiol Rev*, **14**, 447–75.

Adl, S. M., Simpson, A. G., Farmer, M. A., *et al.* (2005). The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J Eukaryot Microbiol*, **52**, 399–451.

Andersson, J. O. (2005). Lateral gene transfer in eukaryotes. *Cell Mol Life Sci*, **62**, 1182–97.

Andersson, J. O. (2006a). Convergent evolution: gene sharing by eukaryotic plant pathogens. *Curr Biol*, **16**, R804–6.

Andersson, J. O. (2006b). Genome evolution of anaerobic protists: metabolic adaptation via gene acquisition. In Katz, L. A., and Bhattacharya, D. (Eds.). *Genomics and evolution of microbial eukaryotes*. Oxford: Oxford University Press.

Andersson, J. O., and Roger, A. J. (2003). Evolution of glutamate dehydrogenase genes: evidence for lateral gene transfer within and between prokaryotes and eukaryotes. *BMC Evol Biol*, **3**, 14.

Andersson, J. O., Sjögren, Å. M., Davis, L. A. M., Embley, T. M., and Roger, A. J. (2003). Phylogenetic analyses of diplomonad genes reveal frequent lateral gene transfers affecting eukaryotes. *Curr Biol*, **13**, 94–104.

Andersson, J. O., Sarchfield, S. W., and Roger, A. J. (2005). Gene transfers from Nanoarchaeota to an ancestor of diplomonads and parabasalids. *Mol Biol Evol*, **22**, 85–90.

Andersson, J. O., Hirt, R. P., Foster, P. G., and Roger, A. J. (2006). Evolution of four gene families with patchy phylogenetic distribution: influx of genes into protist genomes. *BMC Evol Biol*, **6**, 27.

Andersson, J. O., Sjögren, Å. M., Horner, D. S., *et al*. (2007). A genomic survey of the fish parasite *Spironucleus salmonicida* indicates genomic plasticity among diplomonads and significant lateral gene transfer in eukaryote genome evolution. *BMC Genomics*, **8**, 51.

Archibald, J. M., and Keeling, P. J. (2002). Recycled plastids: a 'green movement' in eukaryotic evolution. *Trends Genet*, **18**, 577–84.

Archibald, J. M., Rogers, M. B., Toop, M., Ishida, K., and Keeling, P. J. (2003). Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigelowiella natans*. *Proc Natl Acad Sci USA*, **100**, 7678–83.

Baldauf, S. L. (2003). The deep roots of eukaryotes. *Science*, **300**, 1703–6.

Beiko, R. G., Harlow, T. J., and Ragan, M. A. (2005). Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci USA*, **102**, 14332–7.

Belbahri, L., Calmin, G., Mauch, F., and Andersson, J. O. (2008). Evolution of the cutinase gene family: evidence for lateral gene transfer of a candidate *Phytophthora* virulence factor. *Gene*, **31**, 1–8.

Bergthorsson, U., Adams, K. L., Thomason, B., and Palmer, J. D. (2003). Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature*, **424**, 197–201.

Biagini, G. A., Yarlett, N., Ball, G. E., *et al.* (2003). Bacterial-like energy metabolism in the amitochondriate protozoon *Hexamita inflata*. *Mol Biochem Parasitol*, **128**, 11–9.

Boucher, Y., Douady, C. J., Papke, R. T., *et al.* (2003). Lateral gene transfer and the origins of prokaryotic groups. *Annu Rev Genet*, **37**, 283–328.

Carlton, J. M., Hirt, R. P., Silva, J. C., *et al.* (2007). Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science*, **315**, 207–12.

Cavalier-Smith, T. (2002). The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *Int J Syst Evol Microbiol*, **52**, 297–354.

Ciuffetti, L. M., Tuori, R. P., and Gaventa, J. M. (1997). A single gene encodes a selective toxin causal to the development of tan spot of wheat. *Plant Cell*, **9**, 135–44.

Doolittle, W. F. (1998). You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet*, **14**, 307–11.

Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science*, **284**, 2124–9.

Doolittle, W. F., Boucher, Y., Nesbø, C. L., *et al.* (2003). How big is the iceberg of which organellar genes in nuclear genomes are but the tip? *Philos Trans R Soc Lond B Biol Sci*, **358**, 39–58.

Edwards, J. E., McEwan, N. R., Travis, A. J., and Wallace, J. R. (2004). 16S rRNA library-based analysis of ruminal bacterial diversity. *Antonie van Leeuwenhoek*, **86**, 263–81.

El-Sayed, N. M., Myler, P. J., Blandin, G., *et al.* (2005). Comparative genomics of trypanosomatid parasitic protozoa. *Science*, **309**, 404–9.

Esser, C., Ahmadinejad, N., Wiegand, C., *et al.* (2004). A genome phylogeny for mitochondria among α-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Molecular Biology and Evolution*, **21**, 1643–60.

Figge, R. M., and Cerff, R. (2001). GAPDH gene diversity in spirochetes: a paradigm for genetic promiscuity. *Mol Biol Evol*, **18**, 2240–9.

Filee, J., Siguier, P., and Chandler, M. (2007). I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. *Trends Genet*, **23**, 10–5.

Friesen, T. L., Stukenbrock, E. H., Liu, Z., *et al.* (2006). Emergence of a new disease as a result of interspecific virulence gene transfer. *Nat Genet*, **38**, 953–6.

Garcia-Vallve, S., Romeu, A., and Palau, J. (2000). Horizontal gene transfer of glycosyl hydrolases of the rumen fungi. *Mol Biol Evol*, **17**, 352–61.

Genereux, D. P., and Logsdon, J. M. Jr. (2003). Much ado about bacteria-to-vertebrate lateral gene transfer. *Trends Genet*, **19**, 191–5.

Gile, G. H., Patron, N. J., and Keeling, P. J. (2006). EFL GTPase in cryptomonads and the distribution of EFL and EF-1α in chromalveolates. *Protist*, **157**, 435–44.

Gogarten, J. P. (2003). Gene transfer: gene swapping craze reaches eukaryotes. *Curr Biol*, **13**, R53–4.

Gogarten, J. P., and Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol*, **3**, 679–87.

Gojkovic, Z., Knecht, W., Zameitat, E., *et al.* (2004). Horizontal gene transfer promoted evolution of the ability to propagate under anaerobic conditions in yeasts. *Mol Genet Genomics*, **271**, 387–93.

Hall, C., Brachat, S., and Dietrich, F. S. (2005). Contribution of horizontal gene transfer to the evolution of *Saccharomyces cerevisiae*. *Eukaryot Cell*, **4**, 1102–15.

Harper, J. T., and Keeling, P. J. (2004). Lateral gene transfer and the complex distribution of insertions in eukaryotic enolase. *Gene*, **340**, 227–35.

Huang, J., and Gogarten, J. P. (2006). Ancient horizontal gene transfer can benefit phylogenetic reconstruction. *Trends Genet*, **22**, 361–6.

Huang, J., Mullapudi, N., Lancto, C. A., *et al.* (2004). Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in *Cryptosporidium parvum*. *Genome Biol*, **5**, R88.

Huang, J., Xu, Y., and Gogarten, J. P. (2005). The presence of a haloarchaeal type tyrosyl-tRNA synthetase marks the opisthokonts as monophyletic. *Mol Biol Evol*, **22**, 2142–6.

Jenkins, C., Samudrala, R., Anderson, I., *et al.* (2002). Genes for the cytoskeletal protein tubulin in the bacterial genus *Prosthecobacter*. *Proc Natl Acad Sci USA*, **99**, 17049–54.

Jørgensen, A., and Sterud, E. (2007). Phylogeny of *Spironucleus* (Eopharyngia: Diplomonadida: Hexamitinae). *Protist*, **158**, 157–54.

Keeling, P. J., and Inagaki, Y. (2004). A class of eukaryotic GTPase with a punctate distribution suggesting multiple functional replacements of translation elongation factor 1α. *Proc Natl Acad Sci USA*, **101**, 15380–5.

Kenrick, P., and Crane, P. R. (1997). The origin and early evolution of plants on land. *Nature*, **389**, 33–9.

Kolattukudy, P. E., Rogers, L. M., Li, D., Hwang, C. S., and Flaishman, M. A. (1995). Surface signaling in pathogenesis. *Proc Natl Acad Sci USA*, **92**, 4080–7.

Koonin, E. V. (2003). Horizontal gene transfer: the path to maturity. *Mol Microbiol*, **50**, 725–7.

Kurland, C. G., Canback, B., and Berg, O. G. (2003). Horizontal gene transfer: a critical view. *Proc Natl Acad Sci USA*, **100**, 9658–62.

Latijnhouwers, M., De Wit, P. J., and Govers, F. (2003). Oomycetes and fungi: similar weaponry to attack plants. *Trends Microbiol*, **11**, 462–9.

Lawrence, J. G. (2005). Common themes in the genome strategies of pathogens. *Curr Opin Genet Dev*, **15**, 584–8.

Loftus, B., Anderson, I., Davies, R., *et al.* (2005). The genome of the protist parasite *Entamoeba histolytica*. *Nature*, **433**, 865–8.

Martin, W., Brinkmann, H., Savonna, C., and Cerff, R. (1993). Evidence for a chimeric nature of nuclear genomes: eubacterial origin of eukaryotic glyceraldehyde-3-phosphate dehydrogenase genes. *Proc Natl Acad Sci USA*, **90**, 8692–6.

Martin, W., Rujan, T., Richly, E., *et al.* (2002). Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci USA*, **99**, 12246–51.

Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.

Pál, C., Papp, B., and Lercher, M. J. (2005). Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet*, **37**, 1372–5.

Qian, Q., and Keeling, P. J. (2001). Diplonemid glyceraldehyde-3-phosphate dehydrogenase (GAPDH), and prokaryote-to-eukaryote lateral gene transfer. *Protist*, **152**, 193–201.

Ramesh, M. A., Malik, S. B., and Logsdon, J. M., Jr. (2005). A phylogenomic inventory of meiotic genes; evidence for sex in *Giardia* and an early eukaryotic origin of meiosis. *Curr Biol*, **15**, 185–91.

Reyes-Prieto, A., Hackett, J. D., Soares, M. B., Bonaldo, M. F., and Bhattacharya, D. (2006). Cyanobacterial contribution to algal nuclear genomes is primarily limited to plastid functions. *Curr Biol*, **16**, 2320–5.

Ricard, G., Mcewan, N. R., Dutilh, B. E., *et al.* (2006). Horizontal gene transfer from bacteria to rumen ciliates indicates adaptation to their anaerobic carbohydrates rich environment. *BMC Genomics*, **7**, 22.

Rice, D. W., and Palmer, J. D. (2006). An exceptional horizontal gene transfer in plastids: gene replacement by a distant bacterial paralog and evidence that haptophyte and cryptophyte plastids are sisters. *BMC Biol*, **4**, 31.

Richards, T. A., Hirt, R. P., Williams, B. A., and Embley, T. M. (2003). Horizontal gene transfer and the evolution of parasitic protozoa. *Protist*, **154**, 17–32.

Richards, T. A., Dacks, J. B., Campbell, S. A., *et al.* (2006a). Evolutionary origins of the eukaryotic shikimate pathway: gene fusions, horizontal gene transfer, and endosymbiotic replacements. *Eukaryot Cell*, **5**, 1517–31.

Richards, T. A., Dacks, J. B., Jenkinson, J. M., Thornton, C. R., and Talbot, N. J. (2006b). Evolution of filamentous plant pathogens: gene exchange across eukaryotic kingdoms. *Curr Biol [0960–9822]*, **16**, 1857–64.

Richardson, A. O., and Palmer, J. D. (2006). Horizontal gene transfer in plants. *J Exp Bot*, **58**, 1–9.

Rocap, G., Larimer, F. W., Lamerdin, J., *et al.* (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature*, **424**, 1042–7.

Roxström-Lindquist, K., Palm, D., Reiner, D., Ringqvist, E., and Svärd, S. G. (2006). *Giardia* immunity – an update. *Trends Parasitol*, **22**, 26–31.

Sanders, I. R. (2006). Rapid disease emergence through horizontal gene transfer between eukaryotes. *Trends Ecol Evol*, **21**, 656–8.

Sarti, P., Fiori, P. L., Forte, E., *et al.* (2004). *Trichomonas vaginalis* degrades nitric oxide and expresses a flavorubredoxin-like protein: a new pathogenic mechanism? *Cell Mol Life Sci*, **61**, 618–23.

Schlieper, D., Oliva, M. A., Andreu, J. M., and Lowe, J. (2005). Structure of bacterial tubulin BtubA/B: evidence for horizontal gene transfer. *Proc Natl Acad Sci USA*, **102**, 9170–5.

Shi, N. Q., and Jeffries, T. W. (1998). Anaerobic growth and improved fermentation of *Pichia stipitis* bearing a *URA1* gene from *Saccharomyces cerevisiae*. *Appl Microbiol Biotechnol*, **50**, 339–45.

Simpson, A. G. B., and Roger, A. J. (2004). The real 'kingdoms' of eukaryotes. *Curr Biol*, **14**, R693–6.

Stechmann, A., Baumgartner, M., Silberman, J. D., and Roger, A. J. (2006). The glycolytic pathway of *Trimastix pyriformis* is an evolutionary mosaic. *BMC Evol Biol*, **6**, 101.

Steele, R. E., Hampson, S. E., Stover, N. A., Kibler, D. F., and Bode, H. R. (2004). Probable horizontal transfer of a gene between a protist and a cnidarian. *Curr Biol*, **14**, R298–9.

Sztukowska, M., Bugno, M., Potempa, J., Travis, J., and Kurtz, D. M. Jr. (2002). Role of rubrerythrin in the oxidative stress response of *Porphyromonas gingivalis*. *Mol Microbiol*, **44**, 479–88.

Takishita, K., Ishida, K., and Maruyama, T. (2003). An enigmatic GAPDH gene in the symbiotic dinoflagellate genus *Symbiodinium* and its related species (the order Suessiales): possible lateral gene transfer between two eukaryotic algae, dinoflagellate and euglenophyte. *Protist*, **154**, 443–54.

Temporini, E. D., and VanEtten, H. D. (2004). An analysis of the phylogenetic distribution of the pea pathogenicity genes of *Nectria haematococca* MPVI supports the hypothesis of their origin by horizontal transfer and uncovers a

potentially new pathogen of garden pea: *Neocosmospora boniensis*. *Curr Genet*, **46**, 29–36.

Waller, R. F., Slamovits, C. H., and Keeling, P. J. (2006). Lateral gene transfer of a multigene region from cyanobacteria to dinoflagellates resulting in a novel plastid-targeted fusion protein. *Mol Biol Evol*, **23**, 1437–43.

Watkins, R. F., and Gray, M. W. (2006). The frequency of eubacterium-to-eukaryote lateral gene transfers shows significant cross-taxa variation within amoebozoa. *J Mol Evol*, **63**, 801–14.

Wenzl, P., Wong, L., Kwang-Won, K., and Jefferson, R. A. (2005). A functional screen identifies lateral transfer of β-glucuronidase (*gus*) from bacteria to fungi. *Mol Biol Evol*, **22**, 308–16.

Wolf, Y. I., Aravind, L., Grishin, N. V., and Koonin, E. V. (1999). Evolution of aminoacyl-tRNA synthetases-analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res*, **9**, 689–710.

CHAPTER 13

Lessons in Evolution from Genome Reduction in Endosymbionts

Andrés Moya and Amparo Latorre

13.1. INTRODUCTION

Intracellular bacteria (symbionts and parasites) are characterized by a genome reduction syndrome that, when compared to their free-living relatives, leads us to the conclusion that they are evolving anomalously. Is this right? The notion of anomaly has an anthropocentric connotation, and from such a viewpoint we cannot state that genome reduction is an evolutionary anomaly; likewise we cannot state that parasitic organisms represent a degenerate stage of evolution as compared to their non-parasitic ancestors. By contrast, we can affirm that they represent an anomaly if we are unable to explain their origin and evolution, given there is no suitable theory to explain both the increase in genome size and evolutionary complexity as well as the genome reduction process in endosymbionts. The question is: Do we have such a theory? In the past few years Michel Lynch and collaborators have published a series of papers on this particular issue, trying to integrate the evolution of genome size and concomitantly genome complexity of prokaryotes and eukaryotes into a single theoretical framework following the basic principles of population genetics.

In this chapter, we would like to deal with how basic principles, such as mutation, selection, and effective population size, can give us an acceptable explanation, empirically founded, of the genome reduction process in endosymbiotic bacteria. We propose a model that both explains the huge transformation of endosymbiotic genomes at nucleotide level and accounts for genome reduction. We call this model the size mutational burden model, and it takes advantage of several theoretical studies on organelle population genetics. Although new experiments are needed to verify some predictions of the model, we interpret some results derived from the genome sequencing

of four strains of *Buchnera aphidicola*, the primary endosymbiont of aphids. *B. aphidicola* genome sizes range from 650 kb in *B. aphidicola* from *Acyrthosiphon pisum* (*B. aphidicola* BAp) to 416 kb in *Cinara cedri* (*B. aphidicola* BCc).

13.2. GENOME-SIZE EXPANSION/CONTRACTION

In recent years Lynch and co-workers have developed a theory on genome evolution that, based on the basic principles of population genetics, tries to explain not only the size increase, but also complexity, ranging from prokaryotes to multicellular eukaryotes (Lynch and Conery, 2003; Lynch, 2005, 2006; Lynch *et al.*, 2006). They argue that genome modifications throughout life's history "emerged passively in response to the long-term population-size reductions that accompanied increases in organism size". Surprisingly there is a negative correlation, in logarithmic scale, between genome size and the composite parameter $N_e u$, the product of effective population size by nucleotide mutation rate (Lynch and Conery, 2003). Genome size expansion or contraction is a function of the distribution of insertion/deletion (indels) sizes produced by mutation and the ulterior action of natural selection posing a differential filter to the corresponding sizes. In fact, if natural selection were ineffective, genome size would progressively expand if the rate of insertion exceeded that of deletion, or progressively decline to a minimum level compatible with maintaining gene function if deletions were more abundant (Lynch, 2006).

After discussing the pros and cons of different hypotheses concerning increasing genome size, Lynch (2006) argues in favour of the mutational-burden hypothesis. According to this hypothesis, mutation and selection are not viewed as independent processes, because variation in mutation rate is equivalent to a selective process, as the descendants of some alleles are removed from the population more rapidly than others. In addition, random drift must be considered much more than single evolutionary noise. According to Lynch (2006) "the degree of drift predictably defines the type of pathways that are open to evolution".

In this review we consider whether the evolution of a particular group of bacteria, to be specific, endosymbiotic bacteria from insects, can be explained within the framework of the mutational-burden hypothesis. We then go on to discuss whether this hypothesis gains support as a theory to explain the evolution of genome size and architecture.

Contrary to the general pattern of many prokaryotes, characterized by a relatively high effective population size long-term (as compared to

eukaryotes), endosymbiotic bacteria are apparently subjected to continuous bottlenecking, somehow promoting low effective population size and, probably, putting into doubt certain genome features that are regularly observed in other prokaryotes. However, in order to understand this statement we need to consider how effective population size is affected by mutation rate, because this is a critical composite parameter that passively affects the evolutionary outcome of organisms, at least at genomic scale.

13.3. SOME BACKGROUND INFORMATION ON POPULATION GENETICS

We assume that the vast majority of mutations (from single point mutations to indels of varying size) are mostly deleterious. Let us suppose that any new allele is selected against, with a value of s and a finite population of N_e effective size. It is well known that if s is greater than $1/N_e$, natural selection will remove the new allele from the population effectively, but when s is less than $1/N_e$ the allele will behave neutrally. Moreover, if the allele is completely neutral, the fixation probability, P_f, equals the initial frequency, whereas for a deleterious mutation that probability approximately equals $P_f \approx 2N_e s/(e^{2N_e s} - 1)$ (Lynch, 2006). We must be aware that if $2N_e s$ is small enough, P_f tends to 1, which is a condition of effective neutrality. In contrast, if $2N_e s$ is large enough, then the allele will have a negligible probability of being fixed. Then, it seems that $2N_e s$ is a key parameter to understanding long-term evolution across phylogenetic lineages. However, a major problem concerns the reliable estimation of this parameter. For our purposes, let us consider the deleterious effect of an indel that modifies genome length in n nucleotides. If we replace s with nu, $2N_e s$ becomes $2N_e nu$. The fragment, either an insertion or a deletion, will be fixed provided that $2N_e u$ is much lower than $1/n$. In contrast, the fragment will not be fixed by random drift if $2N_e u$ is much higher than $1/n$.

13.4. ESTIMATING $2N_e u$ IN ENDOSYMBIOTIC BACTERIA

Do we have any idea about this composite parameter in arthropod endosymbionts? Lynch and Conery (2003) estimate $2N_e u$ by measuring levels of silent-site variation among random alleles within a species (i.e., silent site nucleotide diversity). Daubin and Moran (2004) posed some caveats to this procedure, although Lynch and Conery (2004) have defended and argued in favour of their approach.

There are few reports on nucleotide diversity in bacterial endosymbionts. Working with the leucine plasmid of *B. aphidicola*, which belongs to populations of the aphid *Rhopalosiphum padi* distributed worldwide, Martínez-Torres *et al.* (1996) obtained an estimate of 0.0113 mutations per site per year by means of a restriction site analysis. Years later, Funk *et al.* (2001) measured nucleotide diversity of *B. aphidicola* from the aphid *Uroleucon ambrosiae*, based on a few genes that were sequenced from several populations of that aphid, also distributed worldwide. The corresponding estimate was two to three times lower (ranging from 0.0042 to 0.0036) than the one estimated in the former study. Although more studies are needed on this particular estimation, we can state that silent nucleotide diversity in *B. aphidicola* from aphid species ranges between 0.004 and 0.01. As mentioned earlier, in accordance with Lynch and Conery (2003), these values can be used to estimate mutation rate, and also to evaluate the relative power of random drift or natural selection and thus account for the evolution of endosymbiotic bacteria, particularly in terms of genome reduction.

13.5. ESTIMATING EFFECTIVE POPULATION AND MUTATION RATE IN *B. aphidicola*

B. aphidicola, the obligate endosymbiont of aphids, is a Gram-negative bacterium belonging to the gamma-3 Proteobacteria (Munson *et al.*, 1991). The role of endosymbiosis is nutritional, as *B. aphidicola* provides the aphid with essential amino acids that are deficient in its diet (Douglas, 1998). The association between aphids and *B. aphidicola* is very ancient, and the congruence between the phylogenetic trees of hosts and symbionts indicates a sole infection event about 84 to 164 million years ago, followed by the co-evolution of both partners (von Dohlen and Moran, 2000). *B. aphidicola* is maternally inherited by infection of the eggs or embryos and lives in a very closed environment, inside specialized aphid cells called bacteriocytes, which form an organ-like structure called a bacteriome. Wilkinson and Douglas (1998) have shown that an adult from the aphid *Acyrthosiphon pisum* has, on average, 100 bacteriocytes, each containing around 23,500 bacteria. They also reported that the number of bacteria infecting an offspring is reduced, in all likelihood coming from a single bacteriocyte from the mother and, probably, in a number ranging between 50 and 500 (Douglas, 1998). Mira and Moran (2002) also carried out experiments to assess the number of *Buchnera* infecting a new embryo in several aphid species. In *Nasonovia* sp. and in *A. pisum*, they observed quantities as small as 800 and 1,800 bacteria, respectively; whereas, in *U. ambrosiae* the number was around 8,000. All these figures

can be used to obtain some estimates of the effective population size. When real population size fluctuates, the effective population size is the harmonic mean of the different values (Crow and Kimura, 1970) and is influenced to a much greater extent by smaller values than by larger ones. Consequently, we can take the observed number of bacteria at the beginning of embryo development as the lowest estimates. As mentioned earlier, the number of bacteria in adults is several orders of magnitude higher. So we can assume that the effective population size of *B. aphidicola* in aphid species, roughly speaking, has a range of 800–8,000. However, there is another point to consider that, curiously, has not been taken into account and could significantly affect the effective population size. It is well known that *Buchnera* cells, and probably many other bacterial endosymbionts, are polyploid, with a copy number that varies between 50 and 200 during the aphid life cycle (Komaki and Ishikawa, 1999, 2000). According to Wright (1969), the effective size of a k-ploid population of constant size N is kN. Although we need a specific model to deal with genetic variation of bacterial endosymbionts (which we will consider in the next section), we can hypothesize that the effective size of a k-ploid fluctuating population will be close to k times higher for each value used to derive the effective size from the harmonic mean. For instance, if we have two figures for *Buchnera* of 800 and 800,000 at two given moments of its evolutionary history, and an average ploidy per bacterial cell of 150, then $N_e = 239,700$, the harmonic mean of 120,000 and 12×10^7.

If we consider an effective population size of 250,000, and $2N_e u$ ranges between 0.004 and 0.01, then u, the mutation rate per nucleotide site per generation in *Buchnera,* will be 8×10^{-9} and 2×10^{-8}, respectively. Based on a different approach, Brynnel *et al.* (1998) and Ochman *et al.* (1999) have inferred substitution rates in *Buchnera* of $3.9 - 8.2 \times 10^{-9}$, which are very close to those reported here. However, in spite of this agreement, our reasoning may have several weak points. First, the effective population size for this maternally inherited bacterium is, normally, N_f, the effective number of insect host females for the set of populations forming the entire species (Birky *et al.*, 1983). If neutral nucleotide diversity is a reliable estimate of $2N_e u$, N_e must be replaced by N_f and, consequently, our consideration on the copy number of bacterial DNA seems not to affect nucleotide diversity at the population and species levels very much. Second, the aforementioned effective population size seems to be too high as to reduce the effectiveness of both Muller's ratchet and random accumulation of deleterious mutations due to bottlenecking. However, we must bear in mind that we are dealing with the full host species distribution and a long period of the association between *Buchnera* and the aphid host. Consequently, once can assume than

N_f in aphids may be as high as 250,000. With respect to this particular issue, Rispe and Moran (2000) simulated endosymbiont population dynamics and concluded that the accumulation process of deleterious mutations is highly influenced by bacterial effective population size, among other things, and that such an effect declines when the effective population size increases. Third, our approach to estimate the effective population size of *Buchnera* is probably incorrect, because this organism, contrary to the assumption made by Wright (1969), has no constant population size. Consequently, how can we rule out the possibility that bacteria infecting a new embryo have many DNA copies that could, potentially, contribute to genetic diversity at species level? A possible answer relies on the assumption that these copies are genetically identical. Do we have any evidence for such an assumption? An answer can be derived from population genetics of organelles for which many more data are available.

As occurs in insect mitochondrial DNA (mtDNA), effective population size of bacterial endosymbionts, including those inherited maternally, must be affected by the effective size of the corresponding insect host. Then, we can expect mtDNA levels of nucleotide diversity to be comparatively different from the levels found in bacterial endosymbionts if the corresponding mutation rates are different. Take, for instance, mtDNA of aphids. As expected, mtDNA genome size is in the range of animal mtDNA (14–20 kb; Lynch, 2006), this being about 16.4 kb in the particular case of the aphid *R. padi* (Martínez-Torres *et al.*, 1996). In the aforementioned study, mtDNA nucleotide diversity was approached by restriction-site analyses of many aphid strains distributed worldwide. Once mapped, the variable sites mostly corresponded to silent ones. Thus, the mtDNA nucleotide diversity obtained (0.014) should be a good estimation of silent nucleotide diversity. These values ranged, as expected, within the interval of animal mtDNA (0.01–0.07).

13.6. LOOKING FOR SUITABLE POPULATION GENETICS OF BACTERIAL ENDOSYMBIONTS

As already mentioned, Komaki and Ishikawa (1999, 2000) established that chromosome copy number in *Buchnera* ranges from 50 to 200 throughout the aphid life cycle. Previously, Buchner (1965) established a ploidy of 512 for bacteriocytes of cockroach endosymbionts. It can be argued that the ploidy of the endosymbiont, even assuming that most bacterial endosymbionts are polyploids, has a negligible effect on bacterial evolution, because bacterial genetic variation within a single host is null. Do we have any empirical

evidence that is also theoretically well founded? For a proper answer to this question we need a suitable description on the dynamics of bacterial cells inside the host, as well as empirical estimates of bacterial genetic variation within and between single hosts, both within and between populations. Two papers by Birky *et al.* (1983, 1989), considered seminal, modelled bacterial dynamics within the host and also at population levels. The authors were dealing with population genetics of organelles (i.e., mitochondria and chloroplasts) and specifically considered the possibility of genetic variation of organelles within a cell, a situation known as heteroplasmy. It parallels endosymbiotic bacteria if we consider that this bacterium behaves like an organelle. Birky *et al.* (1983) define organelle genetic variability within populations at different levels. More precisely, they denoted K_a, K_b, K_c, and K_d as the probability of sampling two different genes from a single cell of the same individual organism, two different cells of the same individual organism, two different cells from different individuals in a population of organisms, and from different populations of a single species, respectively. When measuring genetic diversity we consider the levels c and d, since it is much harder to measure genetic variability at the level of the individual organism (a and b), something that is particularly relevant when dealing with organelles or, as in our case, with endosymbionts. Birky *et al.* (1983) established the particular conditions under which genetic variation can be observed within individuals. It depends on the ratio between $\bar{t}\,(1/n)$, the mean time to fix or lose a new mutation, and τ, the mean waiting time between mutations in a cell lineage. Both parameters are given by the following two equations:

$$\tau = c/nu \tag{1}$$

$$\bar{t}(1/n) \approx 2\left(\frac{n_e}{n}\right)[\ln(n+1)] \tag{2}$$

where c, u, n, and n_e, for our particular case, are respectively, the number of times bacteriocytes double from the first time until adulthood, the neutral mutation rate per site and per generation, the number of bacterial genomes per bacteriocyte, and the effective number of genomes in a bacteriocyte. Equations (1) and (2) are particularly useful because, depending on the values taken by the four parameters involved, heteroplasmy could be present or not. For instance, for higher mutation rates τ will be lower than for lower ones, and the presence of heteroplasmy will thus be more probable. One expects the same when considering high n_e or n values, because $\bar{t}\,(1/n)$ will be lower than when smaller values of n_e or n are considered. By making use of aforementioned data concerning the number of *Buchnera* in aphids, as well

as the copy chromosome number, the number of doublings, and mutation rates, we have obtained estimates of τ and $\bar{t}\,(1/n)$. Following Wilkinson and Douglas (1998), we assume that bacteriocyte number, throughout the life cycle of an aphid, increases from 1 to 100, so c, the doubling number of bacteriocytes, is about 10. The site mutation rate u is taken as 9×10^{-9} mutations per nucleotide site per replication round (Brynnel *et al.*, 1998; Ochman *et al.*, 1999). For n, which denotes the number of DNA copies of the bacterial chromosome, we consider than a single bacteriocyte has 23,500 copies and 150 DNA copies per bacterium, so $n = 3.5 \times 10^6$. However, we realize that the first figure is probably too high (see the contrasting results reported by Wilkinson and Douglas, 1998, and Mira and Moran, 2000), although the second figure is not (see Komaki and Ishikawa, 1999). Moreover, we should consider that a higher n value implies there is greater likelihood of heteroplasmy within an individual organism. Nevertheless, the most problematic parameter is n_e, the effective number of copies, which will be of the same order as n, at most, and probably much lower. However, as clearly stated by Birky (2001), "the literature reflects a great deal of misunderstanding about the parameter n_e. This is an effective number, which is an unspecified function of the real number of genomes in a cell". As an approach to n_e we have taken into account that both the number of bacteria per bacteriocyte and the DNA copy number for each bacteria vary according to the life cycle. Thus, the harmonic mean was used to estimate n_e under the hypothesis that the true n_e will be more greatly affected both by a low number of bacteria per bacteriocyte and by a low number of DNA copies per bacterium. Consequently, we considered the starting number of DNA copies in a young embryo as 2,500, corresponding to the product of 50 bacteria per bacteriocyte (Douglas, 1998) by 50 DNA copies per bacterium (Komaki and Ishikawa, 1999), and then, on average, the adult aphid will contain the above mentioned 3.5×10^6 DNA bacterial copies per bacteriocyte. Thus, n_e is approximately 5,000, the harmonic mean of 2,500 and 3.5×10^6. τ and $\bar{t}\,(1/n)$ are then 317.46 and 0.04, respectively, according to values of $n = 3.5 \times 10^6$, $n_e = 5,000$, $c = 10$, and $u = 9.0 \times 10^{-9}$. That is to say, it takes a much shorter time for a given mutation to be fixed or lost than for a new one to appear, something that corresponds to the absence of heteroplasmy for an individual organism. In addition, following Birky *et al.* (1983), the mean K_a (i.e., genetic diversity within a single bacteriocyte) taken over a long time will be less than $0.04/316.46 = 0.00013$. This figure is one or two orders of magnitude lower than those previously reported on average neutral genetic variability (i.e., nucleotide diversity) in natural populations of aphids (Martínez-Torres *et al.*, 1986; Funk *et al.* 2001), which are an approach to K_c and K_d (Birky *et al.*,

1983). If n_e approaches n, then the ratio increases to 0.095 and in that case, heteroplasmy should be present and should have an important effect on the evolutionary dynamics of endosymbionts. Birky *et al.* (1983) have shown that one of the consequences of long vegetative segregation is that effective population size for organelle genes can even be higher than for nuclear genes. When transposed to our bacterial organism, we can establish a particular set of conditions where the effective number of bacteria is higher than the effective number of insect hosts.

Let us examine another example. Degnan *et al.* (2004) have estimated that *Blochmannia*, the primary endosymbiont of carpenter ants of the genus *Camponotus*, has a neutral evolutionary rate of 1.3×10^{-7} synonymous substitutions per site per year. If we consider the values of n, n_e, and c for this particular symbiont to be similar to those considered for *Buchnera*, τ and $\bar{t}\,(1/n)$ are 21.98 and 0.04, respectively. Again, the mean K_a taken over a long time will be less than $0.04/21.98 = 0.00182$. This figure is two orders of magnitude lower than the one estimated by our group on the average silent nucleotide diversity in natural populations of *B. floridanus*, primary endosymbiont of the ant *C. floridanus*, sampled from the distribution area of the species (Gómez-Valero *et al.*, in preparation), which is an estimate of K_c and K_d.

The uncertainties about n_e and, to a lesser extent, the other parameters (c, n, and u), can be partly solved if we consider that the different groups working on the evolution of endosymbiotic bacteria have sequenced genes from *Buchnera* and *Blochmannia* by polymerase chain reaction (PCR) on single aphids or ants. We are not aware of a single case in the literature where sequence polymorphism based on a single individual has been reported. On the contrary, the genome sequence of *B. aphidicola* (van Ham *et al.*, 2003) or *B. pennsylvanicus* (Degnan *et al.*, 2005), whose genomic libraries were obtained from mixed individuals of one or more populations, report cases of nucleotide polymorphism.

All these data lead us to the conclusion that our estimates must be close to the real situation. A number of consequences can be drawn from such a conclusion. First, bacteria that reach the offspring, independently of their number and the number of DNA copies, are genetically uniform. Second, contrary to Itoh *et al.* (2002), we reinforce the hypothesis than mutation rate in endosymbiotic lineages does not differ from that in free-living partners. The aforementioned authors argued that increased mutation rate is the major factor contributing to enhanced rate of protein evolution (i.e., nonsynonymous changes) in endosymbiotic bacteria. Our argument, based on a completely different approach, states that such an increase in mutation rate, if present, should be detected within the host, but heteroplasmy has

not been found. Consequently, we are in favour of the relaxation of selection hypothesis (Moran, 1996), which is supported by the observed acceleration of non-synonymous changes as a consequence of the new rich and stable intracellular environment. The model formulated here is compatible with the mutational burden hypothesis (Lynch, 2006) if we can also provide an explanation for the inability of endosymbiotic genomes to increase in size and, according to the empirical evidence, for the systematic reduction in genome size.

13.7. GENOME REDUCTION CAN TAKE PLACE UNDER A SELECTIVE REGIME

What accounts for genome reduction? We should be aware that the change to a new environment where selection is relaxed does not mean that genome reduction will be a natural outcome. On the contrary, we believe that the process of genome reduction is, up to a point, decoupled from the process of nucleotide change. The change towards a high A+T content in bacterial endosymbionts is well documented (i.e., Rispe *et al.*, 2004). Compared with free-living partners, endosymbiont genomes have changed enormously at nucleotide level, although at different rates according to the functional role of the corresponding genes, towards increasing, even exhaustive, levels of A+T. This tendency will promote the appearance of motifs or repeats that will be of importance for indel generation. Moreover, genome size and A+T content in *B. aphidicola* have been shown to be negatively correlated (Gómez-Valero *et al.*, 2004; Pérez-Brocal *et al.*, 2006). Thus, most of the indels should be deletions or, if there are both deletions and insertions, there should be some selective pressure favouring the former over the latter. The possibility of generating indels in bacterial genomes has been analyzed in detail by Rocha (2003). Illegitimate recombination due to slipped mispair or single-strand annealing between close small repeats can be responsible for the mechanistic generation of indels. They are present in *Buchnera* (van Ham *et al.*, 2003) or *Blochmannia* (Degnan *et al.*, 2005) genomes. As already mentioned, the genomic libraries for sequencing the genomes of *B. aphidicola* BBp (van Ham *et al.*, 2003) and *B. pennsylvanicus* (Degnan *et al.*, 2005) were obtained from pooling individuals of one or more populations. After sequencing, polymorphisms were found, some corresponding to indels. There is, then, genome size variation at population and species level (K_c and K_d). Do we have any evidence of genome size variation (K_a and K_b) at the level of a single host? There is no easy answer to this question. Assuming, on the basis of a selective framework, that smaller genomes have an advantage over larger

ones, it is clear that, with respect to a non-reduced genome, one reduced to a single nucleotide will be less advantageous than another one reduced to two nucleotides, and so on. Consequently, we can expect a spectrum of genomes within individuals, within bacteriocytes, and even within bacteria, reaching the homoplasmic state at different rates. For the sake of simplicity, let us consider two examples, with average selection coefficients s of 0.1 and 0.01, which represents a comparative case where the average genome of the first bacteriocyte population has decreased 10 times more than in the second population. The average number of generations until the fixation of such a hypothetical advantageous mutant gene in a given population of bacteriocytes is given by (Kimura and Ohta, 1969):

$$\bar{t}(1/n) = (2/s)\ln(n_e)\,. \tag{3}$$

As $n_e = 5{,}000$, then $\bar{t}(1/n)$ for $s = 0.1$ and 0.01 is 170.3 and 1,703.4 generations, respectively. The time needed to fix a given advantageous mutant is greater than the number c of bacteriocyte generation. Then, this population will remain heteroplasmic within the host. τ is the second factor to consider, the time that elapses between new mutations and, associated to it, u_i, the indel mutation rate. Although there are no direct estimates of this rate, there is well documented evidence of it in DNA from mitochondria and chloroplasts (Birky, 2001). Particularly, there is abundant literature on size variation in insect mtDNA, including aphids. Our group (Martínez-Torres *et al.*, 1996) reported several heteroplasmic regions in mtDNA of the aphid *R. padi*, particularly in the mtDNA A+T-reach region. Illegitimate recombination, as already suggested for the case of bacterial genomes, is also responsible for the generation of indels in those particular A+T-regions. Moreover, it is worth mentioning the similarities between the evolution of this mtDNA region and endosymbiotic bacteria which, in general, also show the highest A+T content of any known bacterial genomes. Assuming similar values of c and n as for the case of nucleotide variability, and that $u_i \geq 10^{-6}$, in the range of those reported for mtDNA A+T rich regions of insects (Rand and Harrison, 1989), $\tau \leq 2.8$ generations. This value points towards mutants of new sizes being generated during within-host bacteriocyte development. This result, together with the length of time that is needed for small advantageous sizes to be fixed, is another factor contributing to size heteroplasmy. Consequently, first we can expect high levels of heteroplasmy, and second that the bacterium invading offspring will also be heteroplasmic. In order to either support or reject all these theoretical considerations, we must urgently carry out suitable experimental approaches to measure the site and size of genetic variability within individual hosts. This issue is of extreme

importance for several reasons. For one, effective population size is affected in that depending on whether or not heteroplasmy is present, effective population size for the bacterial endosymbiont may be greater than or equal to the effective number of host females (Birky *et al.*, 1983). We can then observe situations where there is the absence of site but abundant size heteroplasmy. Muller's ratchet will have a different effect on both types of genetic variation at population or species level.

13.8. MUTATIONAL-BURDEN HYPOTHESIS AND GENOME REDUCTION IN *BUCHNERA*

Is it possible then that a deleterious fragment of length n is lost? Under the mutation burden hypothesis, the condition for the active presence of natural selection requires n to be higher than $1/2N_e u$. Let us suppose the following situation of a population of effective size 250,000 and mutation rate 8×10^{-9}, something that is feasible according to our empirical observations. An effective population size like this has been reported as reliable in the case of insects. Lynch and Conery (2006) summarise that N_e for invertebrates is within the range 10^5 to 10^6. Thus, the effective number of females, taken as the effective number of endosymbionts at population and species level, can be assumed to be in the order mentioned earlier. In this particular case, fragments larger than 250 nt can be removed from the genome by natural selection. In fact, if we consider a situation of much higher effective population sizes or size mutation rates, something that is feasible as already inferred from our considerations on size heteroplasmy of endosymbionts, even fragments of smaller length can be removed by natural selection.

The history of the reductive process in *B. aphidicola* is the most well known from among all the intracellular bacterial endosymbionts studied so far. Four genomes of *B. aphidicola* have been sequenced and the phylogenetic history of this particular endosymbiont has shed some light on the nature of the reductive process. However, there is still some discussion as to whether the mode of genome reduction is taking place through a small number of big deletion events or a large number of small or even single nucleotide deletions (Delmotte *et al.*, 2006). This review has tried to elucidate this issue, using an approach that considers the generation of site and size variation to be two independent, though related, issues. We realize that more experimental work is needed to support some of the questions and ideas outlined here, but the working hypothesis has been formulated and it is just a matter of suitable experimental testing. We agree that the change from a free-living to an intracellular environment contributes to a relaxation of

selection in endosymbiotic bacteria. This particular circumstance, together with low effective population size, guides the evolution of endosymbionts towards accumulating increasingly fewer deleterious mutations with each population. However, this process is slower than expected, probably because the effective population size is higher than expected. Mutation rates at nucleotide level are not really different, on average, from those in free-living partners. It is true that endosymbiotic genomes are losing genes involved in repair, but this loss probably occurred at later stages of symbiotic integration. Moreover, with a higher mutation rate one must expect cases of heteroplasmy, which have not been reported.

We have several reasons for supporting the hypothesis that endosymbiotic genomes are selected for their smaller size, especially when there is abundant gene dispensability and relaxation of selection. First, the high chromosomal copy number displays internal competition among DNA copies when replicating. Although metabolites and enzymes supplied by the host are abundant, probably not enough of them are supplied to avoid such competition. Therefore, smaller genomes will be selected. Second, a high generation rate of within-host size variants, partly due to an increasingly high A+T content in endosymbiotic genomes, must be guaranteed since there is well-known illegitimate recombination among small repeats that generate DNA genomes of different length. It is a key question, then, to evaluate the presence of heteroplasmy for size variation within individual organisms, because the third point is that natural selection will be more effective at species level. This is because the effective population size will be higher than one that considers effective size to be equal to the number of host females, which is expected when there is no heteroplasmy. Finally, it is also important to have a reliable estimate of size mutation rate since it does not necessarily equal site mutation rate and is probably higher.

We end this review by summing up the main features in *B. aphidicola* research. Table 13.1 shows some of them. Genomes are ordered according to size, the smallest corresponding to *B. aphidicola* BCc, from the aphid *Cinara cedri* (Peréz-Brocal *et al.*, 2006). We can speculate as to which factors contribute to the smaller genome size of *B. aphidicola* BCc with respect to the other three *B. aphidicola* that have already been sequenced. As shown in Table 13.1, *B. aphidicola* BCc has only retained four genes involved in repair functions, as compared to the nine in the other *B. aphidicola* strains. In principle, a high mutation rate could be a factor promoting genome size reduction. However, as argued in this work, we do not think that such reduction is primarily due to site mutation rate. As one can observe in Table 13.2, d_s, the average number of synonymous substitutions per site in

Table 13.1. Buchnera *genome size and some features regarding repair genes and effective population sizes of their aphid host*

Aphid	Genome size (bp)	A+T content (%)	Gene number	Repair genes	Intergenic region (IGR) (bp)	Intergenic region A+T content (%)	Average length IGRs (bp)
Acyrthosiphon pisum	652,115	73.7	609	9	73,550	84.7	126.9
Schizaphis graminum	653,001	74.7	597	9	63,578	85.7	113.3
Baizongia pistaciae	618,379	74.7	545	9	101,325	84.6	200.5
Cinara cedri	422,434	79.9	402	4	51,765	91.5	135.8

Table 13.2. *d_n and d_s average values (± standard deviation) among the four* Buchnera *genomes*

Buchnera lineages compared*	Number of genes	d_n	d_s	d_n/d_s
BAp – BSg	349	0.15 ± 0.07	0.72 ± 0.14	0.21
BAp – BBp	348	0.33 ± 0.15	1.05 ± 0.24	0.31
BSg – BBp	343	0.32 ± 0.14	1.01 ± 0.47	0.32
BCc – BAp	355	0.39 ± 0.17	0.86 ± 0.16	0.46
BCc – BSg	350	0.39 ± 0.17	0.83 ± 0.16	0.47
BCc – BBp	350	0.41 ± 0.17	0.86 ± 0.17	0.48

* Abbreviations of the *Buchnera* lineages are as follows: B is from *Buchnera*, and the next two letters correspond to the first letters of the names of the aphid species as described in Table 13.1.

B. aphidicola BCc is no higher than in the other *B. aphidicola* strains. However, we can see that the A+T content is higher in the former than in the other three. Moreover, this value is close to 100% in the intergenic regions (IGR). This might be explained as a consequence of a high size mutation rate, which also varies according to increasing levels of A+T nucleotides, which in turn favours increasing illegitimate recombination events. Regarding the five repair genes lost in *B. aphidicola* BCc, two are involved in correcting, specifically, the effects of that type of illegitimate recombination. On the other hand, as a consequence of a higher relaxation of selection, d_n/d_s, the ratio of non-synonymous to synonymous substitutions (see Table 13.2), is higher in *B. aphidicola* BCc than in the other three strains. This is related not to a plausible explanation of genome reduction, but to the relatively greater loss of functional genes in *B. aphidicola* BCc than in the other *Buchnera* sequenced, which is precisely the reason for the huge difference in genome size. Thus, *B. aphidicola* BCc has lost from 143 to 207 genes as compared to the other three bacteria, whereas there is no reduction in the sizes of the retained intergenic regions or open reading frames (Pérez-Brocal *et al.*, 2006).

Finally, an important issue is related to the total length of the IGR regions in the different *B. aphidicola* strains. As can be observed in Table 13.1, the shortest corresponds to the *B. aphidicola* with the smallest genome, for which we have tried to offer a plausible hypothesis. However, the one corresponding to *B. aphidicola* from *Baizongia pistaciae* possesses the longest IGR despite

the fact that the whole genome is middle-sized when compared to the other three. This is another interesting observation, which is also compatible with the genome-reduction model we are developing here. Apart from the effect of size mutation rate and A+T replacement rate, we must take into account population features of both host and endosymbiont. For instance, in contrast to the other three aphid species, *B. pistaciae* develops in galls, each one derived from a single female. Presumably, this may contribute to reducing the effective endosymbiont population size, at least when compared with the other three. However, mutation rate as derived from d_s (Table 13.2) is the highest of the four *Buchnera,* something that will contribute to the process of genome reduction by increasing repeats of A and/or T that will promote a higher rate of recombination events, which will lead to a higher mutation rate in regard to size.

In short, the genome reduction process is a complex issue, which is affected by several parameters that we have explored here. In order to test the size mutational burden hypothesis of endosymbionts properly, more experiments are needed, especially those concerning within-host genetic variability and population parameters.

ACKNOWLEDGMENTS

Financial support was provided by projects BFM2003–00305 from 'Ministerio de Ciencia y Tecnología', Spain, and ACOMP06/251 from 'Generalitat Valenciana', Spain.

REFERENCES

Birky, C.W. Jr. (2001). The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. *Annu Rev Genet,* **35**, 125–48.

Birky, C.W. Jr., Maruyama, T., and Fuerst, P. (1983). An approach to population and evolutionary genetic theory for genes in mitochondria and chloroplasts, and some results. *Genetics,* **103**, 513–27.

Birky, C. W. Jr., Maruyama, T., and Fuerst, P. (1989). Organelle gene diversity under migration, mutation, and drift: equilibrium expectations, approach to equilibrium, effect of heteroplasmic cells, and comparison to nuclear genes. *Genetics*, **121**, 613–27.

Brynnel, E. U., Kurland, G. C., Moran, N. A., and Andersson, G. C. (1998). Evolutionary rates for *tuf* genes in endosymbionts of aphids. *Mol Biol Evol,* **15**, 574–82.

Buchner, P. (1965). *Endosymbiosis of animals with plant microorganisms*. New York: Interscience.

Crow, J. F., and Kimura, M. (1970). *An introduction to population genetics theory*. New York: Harper and Row.

Daubin, V., and Moran, N. A. (2004). Comment of the "Origins of genome complexity". *Science*, **306**, 978a.

Degnan, P. H., Lazarus, A. B., Brock, C. D., and Wernegreen, J. J. (2004). Host-symbiont stability and fast evolutionary rates in an ant-bacterium association: cospeciation of *Camponotus* species and their endosymbionts, *Candidatus* Blochmannia. *Syst Biol*, **53**, 95–110.

Degnan, P. H., Lazarus, A. B., and Wernegreen, J. J. (2005). Genome sequence of *Blochmannia pennsylvanicus* indicates parallel evolutionary trends among bacterial mutualists of insects. *Genome Res*, **15**, 1023–33.

Delmotte, F., Rispe, C., Schaber, J., Silva, F. J., and Moya, A. (2006). Tempo and mode of early gene loss in endosymbiotic bacteria from insects. *BMC Evol Biol*, **6**, 56.

Douglas, A.E. (1998). Nutritional interactions in insect-microbial symbiosis. *Annu Rev Entomol*, **43**, 17–37.

Funk, D. J., Wernegreen, J. J., and Moran, N. A. (2001). Intraspecific variation in symbiont genomes: bottlenecks and the aphid-*Buchnera* association. *Genetics*, **157**, 477–89.

Gómez-Valero, L., Latorre, A., and Silva, F. J. (2004). The evolutionary fate of nonfunctional DNA in the bacterial endosymbiont *Buchnera aphidicola*. *Mol Biol Evol*, **21**, 2172–81.

Itoh, T., Martin, W., and Nei, M. (2002). Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. *Proc Natl Acad Sci USA*, **99**, 12944–8.

Kimura, M., and Ohta, M. (1969). The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, **61**, 763–71.

Komaki, K., and Ishikawa, H. (1999). Intracellular bacterial symbionts of aphids posses many genomic copies per bacterium. *J Mol Evol*, **48**, 717–22.

Komaki, K., and Ishikawa, H. (2000). Genomic copy number of intracellular bacterial symbionts of aphids varies in response to developmental stage and morph of their host. *Insect Biochem Mol Biol*, **30**, 253–8.

Lynch, M. (2005). The origins of eukaryotic gene structure. *Mol Biol Evol*, **23**, 450–68.

Lynch, M. (2006). Streamlining and simplification of microbial genome architecture. *Annu Rev Microbiol*, **60**, 327–49.

Lynch, M., and Conery, J. S. (2003). The origins of genome complexity. *Science*, **302**, 1401–4.

Lynch, M., and Conery, J. S. (2004). Response to comment on "The origin of genome complexity". *Science*, **306**, 978b.

Lynch, M., Koskella, B., and Schaack, S. (2006). Mutation pressure and the evolution of the organelle genomic architecture. *Science*, **311**, 1727–30.

Martínez-Torres, D., Simon, J. C., Fereres, A., and Moya, A. (1996). Genetic variation in natural populations of the aphid *Rhopalosiphum padi* as revealed by maternally inherited markers. *Mol Ecol*, **5**, 659–70.

Mira, A., and Moran, N. A. (2002). Estimating population size and transmission bottlenecks in maternally transmitted endosymbiotic bacteria. *Microb Ecol*, **44**, 137–43.

Moran, N. A. (1996). Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. *Proc Natl Acad Sci USA*, **93**, 2873–8.

Munson, M. A., Baumann, P., and Kinsey, M. G. (1991). *Buchnera* gen. nov., and *Buchnera aphidicola* sp. nov., a taxon consisting of the mycetocyte-associated, primary endosymbionts of aphids. *Int J Syst Bacteriol*, **41**, 566–8.

Ochmann, H., Elwyn, S., and Moran, N. A. (1999). Calibrating bacterial evolution. *Proc Natl Acad Sci USA*, **96**, 12638–43.

Pérez-Brocal, V., Gil, R., Ramos, S., et al. (2006). A small microbial genome: the end of a long intimate relationship? *Science*, **314**, 312–13.

Rand, D. M., and Harrison, R. G. (1986). Mitochondrial DNA transmission genetics in crickets. *Genetics*, **114**, 955–70.

Rispe, C., and Moran, N. A. (2000). Accumulation of deleterious mutations in endosymbionts: Muller's ratchet with two levels of selection. *Am Naturalist*, **156**, 425–41.

Rispe, C., Delmotte, F., van Ham, R. C. H. J., and Moya, A. (2004). Mutational and selective pressures on codon and amino acid usage in *Buchnera*, endosymbiotic bacteria of aphids. *Genome Res*, **14**, 44–53.

Rocha, E. P. C. (2003). An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: from duplications to genome reduction. *Genome Res*, **13**, 1123–32.

van Ham, R. C. H. J., Kamerbeek, J., Palacios, C., et al. (2003). Reductive genome evolution in *Buchnera aphidicola*. *Proc Natl Acad Sci USA*, **100**, 581–6.

von Dohlen, C. D., and Moran, N. A. (2000). Mitochondrial ribosomal sequences and fossils support a rapid radiation of aphids in the Cretaceous and multiple origins of a complex life cycle. *Biol J Linn Soc*, **71**, 689–717.

Wilkinson, T. L., and Douglas, A. E. (1998). Host cell allometry and regulation of the symbiosis between pea aphids, *Acyrthosiphon pisum*, and bacteria, *Buchnera*. *J Insect Physiol*, **44**, 629–35.

Wright, S. (1969). *Evolution and the genetics of populations*, Vol. 2. Chicago: Chicago University Press.

Index