Springer Protocols

Methods in Molecular Biology 502

Bacteriophages

Methods and Protocols

Volume 2: Molecular and Applied Aspects

Edited by

Martha R. J. Clokie
Andrew M. Kropinski

Humana Press

METHODS IN MOLECULAR BIOLOGY™

Series Editor
John M. Walker
School of Life Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For other titles published in this series, go to
www.springer.com/series/7651

METHODS IN MOLECULAR BIOLOGY™

Bacteriophages

Methods and Protocols, Volume 2 Molecular and Applied Aspects

Edited by

Martha R. J. Clokie
University of Leicester, Leicester, UK

Andrew M. Kropinski
Public Health Agency of Canada, Guelph, Ontario, Canada

Humana Press

Editors
Martha R. J. Clokie
University of Leicester
Leicester, UK
mrjc1@le.ac.uk

Andrew M. Kropinski
Public Health Agency of Canada
Guelph, Ontario, Canada
Andrew_Kropinski@phac-aspc.gc.ca

Series Editor
John M. Walker
University of Hertfordshire
Hatfield, Herts
UK

ISSN 1064-3745
ISBN 978-1-60327-564-4
DOI 10.1007/978-1-60327-565-1

e-ISSN 1940-6029
e-ISBN 978-1-60327-565-1

Library of Congress Control Number: 2008939449

Printed on acid-free paper

springer.com

Preface

We are increasingly aware of the many and varied roles that bacteriophages play in microbial ecology and evolution. The implications of bacteriophage–bacteria interactions range from the evolution of pathogenicity to oceanic carbon cycling. However, working with bacteriophages can be difficult due to their small size and specific bacterial host requirements. Written by top international bacteriophage researchers, these volumes pull together a vast body of knowledge and expertise, including almost forgotten classical methods as well as state-of-the-art molecular techniques. It is designed to be a valuable reference for experienced bacteriophage researchers as well as an accessible introduction to the newcomer to the subject.

The books are designed to be modular and are organised in the order in which one would carry out the work. A wide range of projects can be built from these modules by selecting appropriate chapters from each section. Volume 1s Section 1 concerns the isolation of phages from a range of environments. Sections 2 and 3 describe their morphological and molecular characterisation, and present methods for the investigation of their interaction with bacteria. Volume 2s Sections 1–3 are concerned with bacteriophage genomics, metagenomics, transcriptomics, and proteomics. It concludes with chapters on applied bacteriophage biology (Section 4).

Martha R. J. Clokie
Andrew M. Kropinski

Contents

Contributors

Rafiq Ahmed • *Public Health Agency of Canada, National Laboratory for Enteric Pathogens, National Microbiology Laboratory, Winnipeg, Manitoba, Canada*

Mehdi Arbabi-Ghahroudi • *Institute for Biological Sciences, National Research Council, The Antibody Engineering Group, Ottawa, Ontario, Canada*

Tim Beliën • *Department of Biosystems, Division of Gene Technology, Katholieke Universiteit Leuven, Leuven, Belgium*

Shellie R. Bench • *Ocean Sciences Department, University of California, Santa Cruz, CA, USA*

Jaysheel Bhavsar • *Delaware Biotechnology Institute, University of Delaware, Newark, DE, USA*

Mark Borodovsky • *Wallace H. Coulter Department of Biomedical Engineering and Division of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA*

Pascale Boulanger • *Institut de Biochimie et Biophysique Moléculaire et Cellulaire, Université Paris-Sud, CNRS, Orsay, France*

Tim J. Carver • *Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Sulston Laboratories, Hinxton, Cambridge, UK*

Sherwood R. Casjens • *Division of Cell Biology and Immunology, Department of Pathology, University of Utah Medical School, Salt Lake City, UT, USA*

Ana M. Cerdeño-Tárraga • *Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Sulston Laboratories, Hinxton, Cambridge, UK*

Pieter-Jan Ceyssens • *Department of Biosystems, Division of Gene Technology, Katholieke Universiteit Leuven, Leuven, Belgium*

Irina Chirakadze • *George Eliava Institute of Bacteriophage, Microbiology and Virology (Eliava IBMV), Tbilisi, Georgia*

Jessica L. Clasen • *Department of Earth & Ocean Sciences, University of British Columbia, Vancouver, BC, Canada*

Martha R. J. Clokie • *Lecturer in Microbiology, Department of Infection, Immunity and Inflammation, University of Leicester, Leicester, UK*

Aaron Darling • *University of Wisconsin-Madison, Madison, WI, USA*

Gary Van Domselaar • *National Microbiology Laboratory, Canadian Science Centre for Human and Animal Health, Winnipeg, Manitoba, Canada*

Vincent A. Fischetti • *Laboratory of Bacterial Pathogenesis and Immunology, The Rockefeller University, New York, NY, USA*

Shelley Frost • *Public Health Agency of Canada, Laboratory for Foodborne Diseases, Guelph, Ontario, Canada*

Eddie B. Gilcrease • *Division of Cell Biology and Immunology, Department of Pathology, University of Utah Medical School, Salt Lake City, UT, USA*

Tom Hanson • *Delaware Biotechnology Institute, University of Delaware, Newark, DE, USA*
Kirsten Hertveldt • *Department of Biosystems, Division of Gene Technology, Katholieke Universiteit Leuven, Leuven, Belgium*
Roger P. Johnson • *Public Health Agency of Canada, Laboratory for Foodborne Zoonoses, Guelph, Ontario, Canada*
Andrew M. Kropinski • *Public Health Agency of Canada, Laboratory for Foodborne Diseases, Guelph, Ontario, Canada; Department of Molecular and Cellular Biology, University of Guelph, Guelph, Ontario, Canada; Department of Microbiology and Immunology Queen's University, Kingston, Ontario, Canada*
Rob Lavigne • *Department of Biosystems, Division of Gene Technology, Katholieke Universiteit Leuven, Leuven, Belgium*
Erika Lingohr • *Public Health Agency of Canada, Laboratory for Foodborne Diseases, Guelph, Ontario, Canada*
Alexsandre Lomsadze • *Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA, USA*
Roger MacKenzie • *Institute for Biological Sciences, National Research Council, The Antibody Engineering Group, Ottawa, Ontario, Canada*
Padmanabhan Mahadevan • *Department of Bioinformatics and Computational Biology, George Mason University, Manassas, VA, USA*
Richard A. Manderville • *Department of Chemistry, University of Guelph, Guelph, Ontario, Canada*
David Mead • *Lucigen Corporation, Middleton, WI, USA*
Andrew D. Millard • *Department of Biological Sciences, University of Warwick, Coventry, UK*
Daniel C. Nelson • *University of Maryland Biotechnology Institute, Center for Advanced Research in Biotechnology, Rockville, MD, USA*
Clemens Pausz • *Department of Earth & Ocean Sciences, University of British Columbia, Vancouver, BC, Canada*
Ann Perets • *Public Health Agency of Canada, Laboratory for Foodborne Zoonoses, Office Internationale des Epizooties (OIE) Salmonella Reference Laboratory, Guelph, Ontario, Canada*
Derek John Juan Pickard • *The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK*
J. Robben • *Biomedical Research Institute, Hasselt University, and School of Life Sciences, Transnationale Universiteit Limburg, Diepenbeek, Belgium*
Raymond Schuch • *Laboratory of Bacterial Pathogenesis and Immunology, The Rockefeller University, New York, NY, USA*
Donald Seto • *Department of Bioinformatics and Computational Biology, George Mason University, Manassas, VA, USA*
Jinyu Shan • *Department of Biological Sciences, University of Warwick, Coventry, UK*
Adrian K. Sharma • *Department of Biochemistry and Molecular Biology, Dalhousie University, Nova Scotia, Canada*
Paul Stothard • *Department of Biological Sciences, Biological Sciences Centre, University of Alberta, Edmonton, Alberta, Canada*

ELIZABETH J. SUMMER • *Department Biochemistry and Biophysics, Texas A&M University, College Station, TX, USA*
CURTIS A. SUTTLE • *Departments of Earth and Ocean Sciences, Botany, and Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia, Canada*
JAMSHID TANHA • *Institute for Biological Sciences, National Research Council, The Antibody Engineering Group, Ottawa, Ontario, Canada*
BELA TIWARI • *NERC Environmental Bioinformatics Centre, CEH Oxford, Oxford, UK*
ANDRÉ VILLEGAS • *Public Health Agency of Canada, Laboratory for Foodborne Diseases, Guelph, Ontario, Canada*
GUIDO VOLCKAERT • *Department of Biosystems, Division of Gene Technology, Katholieke Universiteit Leuven, Leuven, Belgium*
DAVID A. WALSH • *Department of Biochemistry and Molecular Biology, Dalhousie University, Nova Scotia, Canada*
DAVID S. WISHART • *Department of Computing Science and Biological Sciences and National Institute for Nanotechnology, University of Alberta, Edmonton, Alberta, Canada*
K. ERIC WOMMACK • *Delaware Biotechnology Institute, University of Delaware, Newark, DE, USA*

Introduction

Andrew M. Kropinski and Martha R. J. Clokie

The discovery of viruses specific to bacteria (referred to variably as bacteriophages, phages, and bacterial viruses in this volume) is credited to an English bacteriologist, Frederick William Twort *(1)* in 1915 and to a French-Canadian scientist, Felix d'Herelle *(2)* in 1917. It is the latter scientist who probably more accurately recognized what he was dealing with and is responsible for naming these agents of bacterial death. He realized these organisms propagated at the expense of bacteria so named them bacteriophages, which translates as bacterial eaters, "phages" coming from the Greek "phagein" meaning "to eat." He is also responsible for recognizing their potential clinical significance *(3)*.

The first golden age of bacteriophage research ran from the 1930s through to the 1970s and resulted in major discoveries such as the identification of DNA as genetic material and the subsequent deciphering of the genetic code and the discovery of messenger RNA, these breakthroughs led to the birth of the new science of Molecular Biology. This work is described in the book by John Cairns et al. *"Phage and the Origins of Molecular Biology" (4)* and on the excellent American Society for Microbiology Division M (Bacteriophage) homepage (http://www.asm.org/division/M/M.html (thanks to Susan Godfrey, Roger Hendrix, Eric Miller). A summary of some of the major discoveries made during this period is detailed in **Table 1** (the authors apologize for the omission of the impact of many eminent phage biologists).

Following on from this golden age was the 1980s and 1990s, where phages and phage-derived products were essential to the major biotechnological revolution that occurred. Recombinant DNA techniques were developed in which phage played a significant part as primary vectors (filamentous phage *(5)*, λ insertional, and replacement vectors *(6)*) or parts of vectors (promoters [expression vector *(7–9)*], packaging signals [cosmids *(10, 11)* and phagemids *(12)*], integrative signals [integrative vectors *(13–15)*], replicons [phagemids *(16)*], and P1-derived vectors *(17)*]. In addition, they contributed a great variety of enzymes are to be employed today's molecular biology laboratory, including integrases, polynucleotide kinases, DNA ligases, DNA polymerases, RNA polymerases, recombinases, single-stranded DNA binding proteins (SSB), endo- and exonucleases, and even methylases and restriction endonucleases *(18)*.

A good indicator for the amount of interest in bacteriophage research is the number of papers published per year that contain the word "bacteriophage" in their title. This rose steadily from 1950 to 1965 and (**Fig. 1**). There was then a sharp burst of phage publications from 1970 to 1975 followed by a precipitous drop in number. This was due to the unfortunate, lack of interest, and funding for phage biology where many eminent scientists gave up working on phages for more lucrative eukaryotic projects.

Recently, due to an increase awareness of their importance, an interest in bacteriophages has been re-kindled, and an insight into the scale of this renewed enthusiasm can be seen from the huge increase in the number of sequenced phage genomes (**Fig. 1**).

Table 1
Significant experimental observation using bacteriophages

Grouping	Discovery	Year & Reference
Plaque assays		Felix d'Herelle 1917 *(50–52)*
Structure and taxonomy	First EM pictures of phages	T.F. Anderson 1942 *(53)*
	CryoEM	1992 *(54–56)*
	Development of modern taxonomic schemes	1962 *(57–59)*
Composition—general	Phages are composed of protein and DNA	1948 *(60)*
	Isolation of: first lipid-containing phage: PM2	1968 *(61)*
Nucleic acids	Genes are made of DNA	1952 *(62)*
	Introns: type I—T4	*(63–65)*
	Inteins	1998 *(66–69)* [http://www.neb.com/inteins.html]
	Genetic code	1961 *(70)*
	Modified bases: T4 (5-hydroxymethylcytosine)	1953 *(71)*
	tRNA-encoding genes: T4	1972 *(72–75)*
	Restriction and modification: λ	
	a) Phenomenon	1953 *(76)*
	b) Mechanism	1962 *(77)*
	Novel genomes:	
	a) Single-stranded (ss) DNA - øX174	1959 *(78)*
	b) Single-stranded (ss) RNA - f2	1961 *(79)*
	c) Segmented double-stranded RNA - ø6	1973 *(80, 81)*
	d) Phage with terminal proteins - ø29	1971 *(82)*
	Sequence of first:	
	a) ssRNA virus	1976 *(83)*
	b) ssDNA virus	1977 *(84)*
Mutation	rII experiments – T4	1955 *(85, 86)*
	T1 resistance in *E.coli*	1943 *(87)*
Lysogeny and integration	Discovery of lysogeny	1934 *(88–90)*
	Isolation of phage λ	E.M. Lederberg 1951 *(91)*
	Induction	1950 *(92)*

(continued)

Table 1 (continued)

Grouping	Discovery	Year & Reference
	Integration:	
	a) Model	1962 *(93)*
	b) Site-specific recombination	1968 *(94, 95)*
	Repression:	
	a) Model	1961 *(96, 97)*
	b) Experimental evidence	1967 *(98, 99)*
	Integration of phage Mu causes host mutations	1963 *(100)*
	Not all temperate phages integrate:	
	a) P1	1951 *(101, 102)*
	b) Linear prophages - N15	N.V. Ravin 1964 *(103)*
	Lysogenic conversion: a) Toxigenicity – *Corynebacterium diphtheriae* phage B	1951 *(104, 105)*
	b) Serotype: *Salmonella* Anatum phage ε15	1955 *(106, 107)*
Genetic exchange - transduction	P22 and *Salmonella*	1952 *(108)*
	P1 and *Escherichia coli*	1955 *(109)*
	Specialized transduction: λ	1957 *(110, 111)*
	Origin of host DNA in P22 transducing particles	1972 *(112)*
Adsorption and injection	a) penetration of capsule	1979 *(113)*
	b) λ & LamB liposomes	1983 *(114)*
	c) T4 and spheroplasts	1983 *(115)*
	d) T7 DNA uptake requires transcription	2001 *(116)*
Intracellular development	One-step growth curve:	
	a) latent period & burst size	1939 *(117)*
	b) burst size from single cells	1945 *(118)*
	c) eclipse phase	1948 *(119, 120)*
	DNA replication:	
	a) DNA ligase	1967 *(121)*
	b) øX174 – rolling circle	1968 *(122, 123)*
	c) T4 – Okazaki fragments	1969 *(124)*
	d) M13 – RNA primers	1972 *(125)*

(continued)

Table 1 (continued)

Grouping	Discovery	Year & Reference
	e) T7 – visualization & formation of concatemers	1972
	General recombination – λ	1961 *(126--129)*
	Transcription:	
	a) mRNA	1956 *(130--133)*
	b) antitermination	1969 *(134)*
	Protein synthesis:	
	a) SDS gels	1969 *(135)*
	b) discontinuous buffer system	1970 *(136)*
	c) slab gel	1973 *(137)*
	d) ribosomal slippage	1993 *(138)*
	Morphogenesis:	
	a) role of chaperonins	1972 *(139--143)*
	b) cross-linked capsid proteins	1995 *(144, 145)*
	c) packaging of φ29 require a small RNA molecule	1987 *(146)*
Phage therapy		F. d'Herelle 1917 *(147)*

From the very steep way in which the slope of the graph of the number of phage genome sequences per year is shooting up, it is quite apparent that we are in a new exponential growth phage of phage research.

There are three main reasons for this renewed interest in bacteriophages. The first is a result of bacterial genome sequencing projects which have revealed that most bacteria are lysogenic for at least one bacteriophage and that phages have played a major role in host genome evolution *(19–24)*. The first project of this kind was the sequencing of the *Haemophilus influenzae* in 1995 which was shown to contain a Mu-like prophage *(25)*. Some bacteria contain many phages in their genomes and pathogenicity is often linked to phage carriage for example the *Streptococcus* group C contain up to 6 phage or prophage-like elements *(26)*. Phages have been shown to encode a range of toxins and gene products that influence their bacterial cells, or even the host in which the bacterium lives (reviewed in *(27)*). An example of the complexity of these interactions can be seen from phages which infect aphid gut bacteria which encode toxins that help bacteria defend the aphid from other invading bacteria *(28)*.

The second reason for the renewed bacteriophage interest is that phage ecologists have shown that soil and water contain between 10 and 100 times more phage particles than bacterial cells, leading to the speculation that the global abundance of phages is probably in the order of 10^{31} *(29)*. Diversity even within phages which infect one bacterial host is also high, for example genomic studies on mycobacterial phages have shown that not only do mycobacterial phage encode genes which are unlike all other genes sequenced thus far, they are also not present in the different phages *(30–35)*. Metagenomic viral studies focusing on the phage in the oceans have also demonstrated the enormous scale of phage

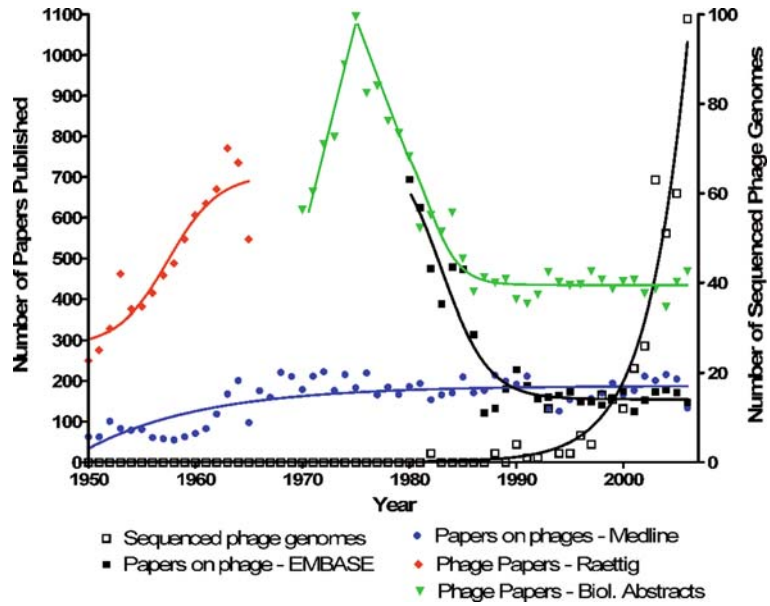


Fig. 1. Publications of bacteriophages and the appearance of phage sequences in GenBank as a function of year. The publication data was derived from four sources: Hansjürgen Raettig for data from 1950 to 1965 *(148, 149)* brought to the editors attention by Hans-Wolfgang Ackermann, and Ovid Medline, EMBASE, and BIOSIS (Biological Abstracts) online literature searches for, respectively, 1950–2006, 1980–2006, and 1969–2006. In the case of Medline and EMBASE, the keyword "bacteriophage" was mapped to subject heading; all subheadings were included; and, the search strategy was focused rather than exploded. With BIOSIS the presence of "bacteriophage or phage" in the title was used to screen scientific articles. NB Although this gives a good indication of phage publications a more detailed analysis is required to tease apart the true number of phage publications, separating those on phage biology from, for example, those on phage typing.

genetic diversity *(29, 36, 37)*. With a raised awareness of phage abundance and diversity has come an appreciation for the consequence of phage action in influencing bacterial population dynamics and evolution and in maintaining essential biogeochemical cycles such as carbon cycling *(38)*. Furthermore, recent genomic and transcriptomic studies have illustrated the extent of interlinked metabolisms of phage and host during a lytic infection for example with bacteriophages which infect cyanobacteria encoding and expressing key photosynthesis gene *(39–42)*.

Finally, but very importantly in terms of phage research are the concerns of the public, governmental healthcare agencies and physicians that something must be done about the growing problem of antimicrobial resistance. This awareness is accompanied by the belated realization that phage therapy, which has been kept alive by the efforts of Eastern European scientists, offers a viable alternative to antibiotic therapy. In Canada, for example, it is now realized that much of the expertise in phage biology has disappeared as a result of retirements and the death of members of the phage community of scientists. The lack of "capacity issues" (i.e., knowledgeable young scientists) has results in the Canadian Institutes for Health Research issuing a call for research grants which will address the potential for using phage as a therapeutic agents. Similarly the same awareness in Europe and the United States has resulted in the number of new bacteriophage research groups increasing and the interest and attendance at bacteriophage conferences is increasing annually. What is apparent, however, is that for bacteriophages to be used therapeutically in countries such

as Europe, Canada and the US, we must properly understand the biology of the interaction between the phage and the bacterial pathogen. We are fortunate to be practicing phage biology in this exciting time where such experimentation is possible.

These volumes are designed to provide the amateur or professional with a step-by-step approach to many of the standard protocols in working with bacteriophages. We include both classical protocols which have been collected before they are forgotten and have to be re-invented, and also state-of-the-art protocols which use many of the latest molecular tools with which to study bacteriophages. It is a complete piece of biology and should take the new comer to bacteriophages from isolating these organisms to characterizing them at every level. It should also equip the experienced phage practitioner wishing to branch out into a new area of phage biology.

Unfortunately with time and space restrictions, it is not possible to be fully comprehensive. When we approached one scientist to contribute to this book he/she replied, "That's microbial archeology. I no longer have access to those laboratory research manuals." For similar reasons, phage immunoelectron microscopy is not covered, nor are the uses of maxi- *(43, 44)* or minicells *(45–49)* for studying phage gene expression.

We are indebted to our authors who have kindly shared their years of experience to make this volume possible. They represent a truly multidisciplinary assemblage of scientists with a huge combined skill set. *Bacteriophages: Methods and Protocols* is a complete piece of biology, laid out in seven sections. Volume 1, Section 1 deals with methods of isolating bacteriophage (and archeophage) from a range of soil or aquatic environments using direct isolation and enrichment approaches. Volume 1, Section 2 covers the characterization of bacteriophages based upon their ability to form plaques, or direct enumeration by fluorescent microscopy or flow cytometry. There is also a chapter on electron microscopy and a further one on classical phage taxonomy. The characterization of host range, adsorption and receptor interaction, and models of plaque development are also considered here. Finally there is a chapter on how to maintain phage stocks once you have them.

Bacteriophage-host interactions (Volume 1, Section 3) includes the construction of mutants using chemical mutagenesis or by recombineering, studies on lysogens, and transduction by temperate and lytic phages. A full scope of genomics is covered in Volume 2, Section 1, from DNA isolation and characterization (PFGE, base composition), through library construction, sequencing, annotation (termini, genes, promoters, terminators) and phylogenetics. Volume 2, Section 2 describes concentrates on transcriptomics and proteomic approaches. These include: mRNA extraction during host infection, quantification of mRNA using real time PCR, and microarray construction. Isolation-independent methods of characterizing phage communities are described in Volume 2, Section 3. Volume 2, Section 4 describes the applied aspects of bacteriophage biology including phage typing, the isolation of lysins, and general and antibody phage display. There is also a final chapter to describe some online resources for phage workers.

To conclude, we hope that you find this book useful and inspiring, and we look forward to the next golden age of phage research.

References

1. Twort, F. W. (1915) *Lancet* **189**, 1241–1243.
2. d'Herelle, F. (1917) *Comptes rendus Académie Sciences* **165**, 373–375.
3. d'Herelle, F. (1926) *The bacteriophage and its behavior* (The Williams & Wilkins Company, Baltimore, MD).

4. Cairns, J., Stent, G. S., & Watson, J. D. (1966) *Phage and the Origins of Molecular Biology*. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.).
5. Smith, G. P. (1988) in *Vectors: A survey of molecular cloning vectors and their uses*, eds. Rodrigues, L. C. & Denhardt, D. T. (Butterworths, Boston), pp. 61–83.
6. Sorge, J. A. (1988) in *Vectors: A survey of molecular cloning vectors and their uses*, eds. Rodriguez, R. L. & Denhardt, D. T. (Butterworths, Boston), pp. 43–60.
7. Denhardt, D. T. & Colasanti, J. (1988) in *Vectors: A survey of molecular cloning vectors and their uses*, eds. Rodrigues, L. C. & Denhardt, D. T. (Butterworths, Boston), pp. 179–203.
8. Brosius, J. (1988) in *Vectors: A survey of molecular cloning vectors and their uses*, eds. Rodrigues, L. C. & Denhardt, D. T. (Butterworths, Boston), pp. 205–225.
9. Mackie, G. A. (1988) in *Vectors: A survey of molecular cloning vectors and their uses*, eds. Rodrigues, L. C. & Denhardt, D. T. (Butterworths, Boston), pp. 253–267.
10. Collins, J. & Bruning, H. J. (1978) *Gene* **4**, 85–107.
11. Hohn, B., Koukolíková-Nicola, Z., Lindenmaier, W., & Collins, J. (1988) in *Vectors: A survey of molecular cloning vectors and their uses*, eds. Rodrigues, L. C. & Denhardt, D. T. (Butterworths, Boston), pp. 113–127.
12. Melnikov, A. A., Tchernov, A. P., Fodor, I., & Bayev, A. A. (1984) *Gene* **28**, 29–35.
13. Hermesz, E., Olasz, F., Dorgai, L., & Orosz, L. (1992) *Gene* **119**, 9–15.
14. Kirsanov, N. B., Mar'ina, O. V., & Ianenko, A. S. (1993) *Genetika* **29**, 1806–1810.
15. Soldatova, L. I., Sladkova, I. A., & Orekhov, A. V. (1994) *Antibiotiki i Khimioterapiia* **39**, 3–7.
16. Mead, D. A. & Kemper, B. (1988) in *Vectors: A survey of molecular cloning vectors and their uses*, eds. Rodrigues, L. C. & Denhardt, D. T. (Butterworths, Boston), pp. 85–102.
17. Cesareni, G. (1988) in *Vectors: A survey of molecular cloning vectors and their uses*, eds. Rodrigues, L. C. & Denhardt, D. T. (Butterworths, Boston), pp. 103–111.
18. Roberts, R. J., Vincze, T., Posfai, J., & Macelis, D. (2003) *Nucleic Acids Research* **31**, 418–420.
19. Brabban, A. D., Hite, E., & Callaway, T. R. (2005) *Foodborne Pathogens and Disease* **2**, 287–303.
20. Boyd, E. F., Davis, B. M., & Hochhut, B. (2001) *Trends in Microbiology* **9**, 137–144.
21. Brussow, H., Canchaya, C., & Hardt, W. D. (2004) *Microbiology & Molecular Biology Reviews* **68**, 560–602.
22. Canchaya, C., Fournous, G., & Brussow, H. (2004) *Molecular Microbiology* **53**, 9–18.
23. Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M. L., Brussow, H., Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M. L., & Brussow, H. (2003) *Current Opinion in Microbiology* **6**, 417–424.
24. Canchaya, C., Proux, C., Fournous, G., Bruttin, A., & Brüssow, H. (2003) *Microbiology & Molecular Biology Reviews* **67**, 238–276.
25. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M. et al. (1995) *Science* **269**, 496–512.
26. Banks, D. J., Beres, S. B., & Musser, J. (2005) in *Phages: Their Role in Bacterial Pathogenesis and Biotechnology*, eds. Waldor, M. K., Friedman, D. I., & Adhya, S. L. (ASM Press, Washington, D.C.), pp. 319–334.
27. Wagner, P. L., Waldor, M. K., Wagner, P. L., & Waldor, M. K. (2002) *Infection & Immunity* **70**, 3985–3993.
28. Moran, N. A., Degnan, P. H., Santos, S. R., Dunbar, H. E., Ochman, H., Moran, N. A., Degnan, P. H., Santos, S. R., Dunbar, H. E., & Ochman, H. (2005) *Proceedings of the National Academy of Sciences of the United States of America* **102**, 16919–16926.
29. Rohwer, F. (2003) *Cell* **113**, 141.
30. Lee, S., Kriakov, J., Vilcheze, C., Dai, Z., Hatfull, G. F., & Jacobs, W. R., Jr. (2004) *FEMS Microbiology Letters* **241**, 271–276.
31. Hatfull, G. F. & Sarkis, G. J. (1993) *Molecular Microbiology* **7**, 395–405.
32. Ford, M. E., Sardis, G. J., Belanger, A. E., Hendrix, R. W., & Hatfull, G. F. (1998) *Journal of Molecular Biology* **279**, 143–164.
33. Ford, M. E., Stenstrom, C., Hendrix, R. W., & Hatfull, G. F. (1998) *Tubercle & Lung Disease* **79**, 63–73.
34. Hatfull, G. F., Pedulla, M. L., Jacobs-Sera, D., Cichon, P. M., Foley, A., Ford, M. E., Gonda, R. M., Houtz, J. M., Hryckowian, A. J., Kelchner, V. A. et al. (2006) *PLoS Genetics* **2**, e92.
35. Pedulla, M. L., Ford, M. E., Houtz, J. M., Karthikeyan, T., Wadsworth, C., Lewis, J. A., Jacobs-Sera, D., Falbo, J., Gross, J., Pannunzio, N. R. et al. (2003) *Cell* **113**, 171–182.
36. Breitbart, M., Felts, B., Kelley, S., Mahaffy, J. M., Nulton, J., Salamon, P., & Rohwer, F. (2004) *Proceedings of the Royal Society of*

London - Series B: Biological Sciences **271**, 565–574.
37. Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., Azam, F., & Rohwer, F. (2002) *Proceedings of the National Academy of Sciences of the United States of America* **99**, 14250–14255.
38. Suttle, C. A. & Suttle, C. A. (2005) *Nature* **437**, 356–361.
39. Mann, N. H., Clokie, M. R., Millard, A., Cook, A., Wilson, W. H., Wheatley, P. J., Letarov, A., & Krisch, H. M. (2005) *Journal of Bacteriology* **187**, 3188–3200.
40. Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M., Chisholm, S. W., Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M., & Chisholm, S. W. (2005) *Nature* **438**, 86–89.
41. Sullivan, M. B., Coleman, M. L., Weigele, P., Rohwer, F., Chisholm, S. W., Sullivan, M. B., Coleman, M. L., Weigele, P., Rohwer, F., & Chisholm, S. W. (2005) *Plos Biology* **3**, e144.
42. Clokie, M. R., Shan, J., Bailey, S., Jia, Y., Krisch, H. M., West, S., Mann, N. H., Clokie, M. R. J., Shan, J., Bailey, S. et al. (2006) *Environmental Microbiology* **8**, 827–835.
43. Mayo, O., Hernandez-Chico, C., & Moreno, F. (1988) *Journal of Bacteriology* **170**, 2414–2417.
44. East, A. K. & Errington, J. (1989) *Gene* **81**, 35–43.
45. Reeve, J. N. (1977) *Molecular & General Genetics* **158**, 73–79.
46. Reeve, J. N. & Cornett, J. B. (1975) *Journal of Virology* **15**, 1308–1316.
47. Mertens, G., Amann, E., Reeve, J. N., Mertens, G., Amann, E., & Reeve, J. N. (1979) *Molecular & General Genetics* **172**, 271–279.
48. Garcia, J. A. & Salas, M. (1980) *Molecular & General Genetics* **180**, 539–545.
49. Reeve, J. (1979) *Methods in Enzymology* **68**, 493–503.
50. Gratia, J.-P. (2000) *Genetics* **156**, 471–476.
51. Gratia, A. (1936) *Annales de l'Institut Pasteur* **57**, 652–676.
52. Adams, M. D. (1959) *Bacteriophages* (Interscience Publishers, Inc., New York).
53. Luria, S. E., Delbrück, M., & Anderson, T. F. (1943) *Journal of Bacteriology* **46**, 57–67.
54. Jiang, W., Chang, J., Jakana, J., Weigele, P., King, J., Chiu, W., Jiang, W., Chang, J., Jakana, J., Weigele, P. et al. (2006) *Nature* **439**, 612–616.
55. Wikoff, W. R., Conway, J. F., Tang, J., Lee, K. K., Gan, L., Cheng, N., Duda, R. L., Hendrix, R. W., Steven, A. C., & Johnson, J. E. (2006) *Journal of Structural Biology* **153**, 300–306.
56. Dokland, T., Lindqvist, B. H., & Fuller, S. D. (1992) *EMBO Journal* **11**, 839–846.
57. Bradley, D. E. & Bradley, D. E. (1966) *Journal of General Microbiology* **44**, 383–391.
58. Bradley, D. E. (1967) *Journal of Bacteriology* **31**, 230–314.
59. Lwoff, A., Horne, R. W., & Tournier, P. (1962) *Cold Spring Harbor Symposia on Quantitative Biology* **27**, 51–62.
60. Putman, F. W., Kozloff, L. M., & Evans, E. A. J. (1948) *Federation Proceedings* **7**, 179.
61. Espejo, R. T. & Canelo, E. S. (1968) *Virology* **34**, 738–747.
62. Hershey, A. D. & Chase, M. (1952) *Journal of General Physiology* **1**, 39–56.
63. Chu, F. K., Maley, G. F., Maley, F., & Belfort, M. (1984) *Proceedings of the National Academy of Sciences of the United States of America* **81**, 3049–3053.
64. Chu, F. K., Maley, G. F., Maley, F., & Belfort, M. (1984) *Federation Proceedings* **43**.
65. Chu, F. K., Maley, G. F., Belfort, M., & Maley, F. (1985) *Federation Proceedings* **44**.
66. Lazarevic, V., Soldo, B., Düsterhöft, A., Hilbert, H., Mauël, C., & Karamata, D. (1998) *Proceedings of the National Academy of Sciences of the United States of America* **95**, 1692–1697.
67. Lazarevic, V. (2001) *Nucleic Acids Research* **29**, 3212–3218.
68. Perler, F. B. (2006) *Nucleic Acids Research* **30**, 383–384.
69. Derbyshire, V. & Belfort, M. (1998) *Proceedings of the National Academy of Sciences of the United States of America*. **95**, 17.
70. Crick, F. H. C., Barnett, L., Brenner, S., & Watts-Tobin, R. J. (1961) *Nature* **192**, 1227–1232.
71. Wyatt, G. R. & Cohen, S. S. (1953) *Annales de l'Institut Pasteur* **84**, 143–146.
72. Guthrie, C., McClain, W. H., Guthrie, C., & McClain, W. H. (1973) *Journal of Molecular Biology* **81**, 137–155.
73. McClain, W. H., Guthrie, C., Barrell, B. G., McClain, W. H., Guthrie, C., & Barrell, B. G. (1972) *Proceedings of the National Academy of Sciences of the United States of America* **69**, 3703–3707.
74. Guthrie, C., Seidman, J. G., Altman, S., Barrell, B. G., Smith, J. D., McClain, W. H., Guthrie, C., Seidman, J. G., Altman, S., Barrell, B. G. et al. (1973) *Nature - New Biology* **246**, 6–11.
75. McClain, W. H., Guthrie, C., Barrell, B. G., McClain, W. H., Guthrie, C., & Barrell, B.

G. (1973) *Journal of Molecular Biology* **81**, 157–171.
76. Bertani, G. & Weigle, J. J. (1953) *Journal of Bacteriology* **65**, 113–121.
77. Arber, W. & Dussoix, D. (1962) *Journal of Molecular Biology* **5**, 18–36.
78. Sinsheimer, R. L. (1959) *Brookhaven Symposia in Biology* **No 12**, 27–34.
79. Loeb, T. & Zinder, N. D. (1961) *Proceedings of the National Academy of Sciences of the United States of America* **47**, 282–289.
80. Semancik, J. S., Vidaver, A. K., & Van Etten, J. L. (1973) *Journal of Molecular Biology* **78**, 617–625.
81. Vidaver, A. K., Koski, R. K., & Van Etten, J. L. (1973) *Journal of Virology* **11**, 799–805.
82. Ortin, J., Viñuela, E., Salas, M., & Vásquez, C. (1971) *Nature New Biology* **234**, 275–277.
83. Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min, J. W., Molemans, F., Raeymaekers, A., Van den, B. A. et al. (1976) *Nature* **260**, 500–507.
84. Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., Hutchison, C. A., Slocombe, P. M., & Smith, M. (1977) *Nature* **265**, 687–695.
85. Benzer, S. (1955) *Proceedings of the National Academy of Sciences of the United States of America* **41**, 344–354.
86. Benzer, S. & Benzer, S. (1956) *Brookhaven Symposia in Biology* 3–5.
87. Luria, S. & Delbrück, M. (1943) *Genetics* **2**, 491–511.
88. Wollman, E. (1934) *Bulletin de L'Institut Pasteur* **32**, 945–955.
89. Bertani, G. (2004) *Journal of Bacteriology* **186**, 595–600.
90. Burnet, F. M. (1934) *Biological Review* **9**, 332–350.
91. Lederberg, E. M. & Lederberg, J. (1953) *Genetics* **38**, 51–64.
92. Lwoff, A., Siminovitch, L., & Kjeldgaard, N. (1950) *Annales de l'Institut Pasteur* **79**, 815–859.
93. Campbell, A. (1962) *Advances in Genetics* **11**, 101–145.
94. Signer, E. R. & Weil, J. (1968) *Cold Spring Harbor Symposia on Quantitative Biology* **33**, 715–719.
95. Echols, H., Gingery, R., & Moore, L. (1968) *Journal of Molecular Biology* **34**, 251–260.
96. Jacob, F. & Monod, J. (1961) *Journal of Molecular Biology* **3**, 318–356.
97. Monod, J. & Jacob, F. (1961) *Cold Spring Harbor Symposia on Quantitative Biology* **26**, 389–401.
98. Ptashne, M. (1967) *Nature* **214**, 232–234.
99. Ptashne, M. (1967) *Proceedings of the National Academy of Sciences of the United States of America* **57**, 306–313.
100. Taylor, A. L. (1963) *Proceedings of the National Academy of Sciences of the United States of America* **50**, 1043–1051.
101. Bertani, L. E. (1951) *Journal of Bacteriology* **62**, 293–300.
102. Ikeda, H. & Tomizawa, J. I. (1968) *Cold Spring Harbor Symposia on Quantitative Biology* **33**, 791–798.
103. Ravin, N. V. (2006) in *The Bacteriophages*, ed. Calendar, R. (Oxford University Press, New York, NY), pp. 448–456.
104. Freeman, V. J. & Morse, I. U. (1952) *Journal of Bacteriology* **63**, 407–414.
105. Freeman, V. J. (1951) *Journal of Bacteriology* **61**, 675–688.
106. Uetake, H., Nakagawa, T., & Akiba, T. (1955) *Journal of Bacteriology* **69**, 571–579.
107. Uetake, H., Luria, S. E., & Burrous, J. W. (1958) *Virology* **5**, 68–91.
108. Zinder, N. D. & Lederberg, J. (1952) *Journal of Bacteriology* **64**, 679.
109. Lennox, E. S. (1955) *Virology* **1**, 190–206.
110. Morse, M. L., Lederberg, E. M., & Lederberg, J. (2007) *Genetics* **41**, 142–156.
111. Weigle, J. (1957) *Virology* **4**, 14–25.
112. Ebel-Tsipis, J., Botstein, D., & Fox, M. S. (1972) *Journal of Molecular Biology* **71**, 433–448.
113. Bayer, M. E., Thurow, H., & Bayer, M. H. (1979) *Virology* **94**, 95–118.
114. Roessner, C. A., Struck, D. K., & Ihler, G. M. (1983) *Journal of Biological Chemistry* **258**, 643–648.
115. Furukawa, H., Kuroiwa, T., & Mizushima, S. (1983) *Journal of Bacteriology* **154**, 938–945.
116. Molineux, I. J. (2001) *Molecular Microbiology* **40**, 1–8.
117. Ellis, E. L. & Delbrück, M. (2007) *Journal of General Physiology* **22**, 365–384.
118. Delbrück, M. (1945) *Journal of Bacteriology* **50**, 131–135.
119. Doermann, A. H. & Dissosway, C. (1948) *Year Book Carnegie Institute of Washington* **48**, 170–176.
120. Doermann, A. H. (1952) *Journal of General Physiology* **35**, 645–656.
121. Gellert, M. (1967) *Proceedings of the National Academy of Sciences of the United States of America* **57**, 148–155.
122. Dressler, D. (1970) *Proceedings of the National Academy of Sciences of the United States of America* **67**, 1934–1942.
123. Gilbert, W. & Dressler, D. (1968) *Cold Spring Harbor Symposia on Quantitative Biology* **33**, 473–484.

124. Okazaki, T. & Okazaki, R. (1969) *Proceedings of the National Academy of Sciences of the United States of America* **64**, 1242–1248.
125. Wickner, W., Brutlag, D., Schekman, R., & Kornberg, A. (1972) *Proceedings of the National Academy of Sciences of the United States of America* **69**, 965–969.
126. Meselson, M. & Weigle, J. J. (1961) *Proceedings of the National Academy of Sciences of the United States of America* **47**, 857–868.
127. Wolfson, J., Dressler, D., & Magazin, M. (1972) *Proceedings of the National Academy of Sciences of the United States of America* **69**, 499–504.
128. Wolfson, J., Dressler, D., & Magazin, M. (1972) *Proceedings of the National Academy of Sciences of the United States of America* **69**, 499–504.
129. Watson, J. D. (1972) *Nature New Biology* **239**, 197–201.
130. Brenner, S., Jacob, F., & Meselson, M. (1961) *Nature* **190**, 576–581.
131. Astrachan, L. & Volkin, E. (1959) *Biochimica et Biophysica Acta* **32**, 449–456.
132. Volkin, E. & Astrachan, L. (1956) *Virology* **2**, 433–437.
133. Volkin, E., Astrachan, L., & Countryman, J. L. (1958) *Virology* **6**, 545–555.
134. Roberts, J. W. (1969) *Nature* **224**, 1168–1174.
135. Weber, K. & Osborn, M. (1969) *Journal of Biological Chemistry* **244**, 4406–4412.
136. Laemmli, U. K. (1970) *Nature* **227**, 680–685.
137. Dunn, J. J. & Studier, F. W. (1973) *Proceedings of the National Academy of Sciences of the United States of America* **79**, 237–248.
138. Levin, M. E., Hendrix, R. W., & Casjens, S. R. (1993) *Journal of Molecular Biology* **234**, 124–139.
139. Coppo, A., Manzi, A., Pulitzer, J. F., & Takahashi, H. (1973) *Journal of Molecular Biology* **76**, 61–87.
140. Georgopoulos, C. P., Hendrix, R. W., Casjens, S. R., & Kaiser, A. D. (1973) *Journal of Molecular Biology* **76**, 45–60.
141. Revel, R. H., Stitt, B. L., Lielausis, I., & Wood, W. B. (1980) *Journal of Virology* **33**, 366–376.
142. Sternberg, N. (1973) *Journal of Molecular Biology* **76**, 25–44.
143. Takano, T. & Kakefuda, T. (1972) *Nature New Biology* **239**, 34–37.
144. Duda, R. L., Martincic, K., Xie, Z., & Hendrix, R. W. (1995) *FEMS Microbiology Reviews* **17**, 41–46.
145. Duda, R. L. (1998) *Cell* **94**, 55–60.
146. Guo, P., Erickson, S., & Anderson, D. (1987) *Science* **236**, 690–694.
147. Dubos, R. J., Straus, J. H., & Pierce, C. (1943) *Journal of Experimental Medicine* **20**, 161–168.
148. Raettig, H. (1958) *Bacteriophagie, 1917 bis 1956; Zugleich en Vorschlag zur Dokumentation Wissenschaftlichen Literatur*. (G. Fisher, Stuttgart).
149. Raettig, H. (1967) *Bacteriophagie 1957–1965* (G. Fisher, Stuttgart).

Section I

Bacteriophage Genomics

Chapter 1

Preparation of Bacteriophage Lysates and Pure DNA

Derek John Juan Pickard

Abstract

Preparation of pure bacteriophage DNA used to rely on using CsCl gradients to give high purity or methods that yielded DNA that was either of low recovery or subject to significant genomic contamination. Recently though, new methods have come along that allow the purification of DNA from plate lysates that are not only capable of high yield but also, for all intents and purposes, free of genomic contamination (i.e. no visible genomic contamination on restriction analysis or when used for bacteriophage sequencing).

This protocol that form the basis of this short section can be used to prepare bacteriophage DNA from one or two 9 cm L-agar plates. For these preps, the use of agarose in the top agar is recommended to avoid any restriction inhibitors that may be present in some agar preparations.

Key words: DNA isolation from phage, bacteriophage lysates, Phase-lock gel purification, pure phage DNA.

1 Introduction

The purification of bacteriophage DNA for restriction digest analysis and other procedures such as sequencing can be carried out using either CsCl gradient purification of phage particles and subsequent lysis to release the DNA or selective purification of phage DNA away from bacterial DNA contamination. The latter techniques take advantage of the fact that the phage DNA is protected by the intact phage within the capsid, while the genomic DNA of the bacteria can be separately degraded by DNase I enzyme. This enzyme, after allowing a short time for digestion of genomic DNA to occur, can be quickly inactivated by addition of proteinase K. The phage coat protecting the phage DNA can then be completely degraded by addition of phenol–chloroform. The phage DNA can then be purified.

Martha R. J. Clokie, Andrew M. Kropinski (eds.), *Bacteriophages: Methods and Protocols, Volume 2: Molecular and Applied Aspects, vol. 502*, © 2009 Humana Press, a part of Springer Science+Business Media
DOI 10.1007/978-1-60327-565-1_1 Springerprotocols.com

2 Materials

For each method that will be described, a number of reagents and other materials will be required. These are detailed below.

2.1 METHOD ONE—Using the Promega Wizard Lambda Preps DNA Purification System Without a Vacuum Source (Product Number: A7290)

1. 5 ml Luer-Lok syringes (Sigma)
2. Proteinase K solution –20 mg/ml (Roche Cat. No. 03 115 828 001)
3. 80% Isopropanol
4. E buffer or similar.

2.2 Modified Version of the Wizard Lambda Preps DNA Purification System Using the Promega Vac-Man Laboratory Manifold

1. 20 mg/ml Proteinase K solution (Roche Cat. No. 03 115 828 001)
2. 80% Isopropanol
3. TE or E buffer
4. Promega Vac-Man Laboratory Vacuum Manifold (either the 20 sample capacity version with Cat. No. A7231 or the Vac-Man Jr. Laboratory Vacuum Manifold that has a two-sample capacity, with Cat. No. A7660)

2.3 Using Phase Lock Gel (Phenol–Chloroform Method) 1.5 ml Eppendorfs or 15 ml Falcons

1. DNase I (Roche Cat. No. 11 284 932 001)
2. RNase A (Roche Cat. No. 10 109 142 001)
3. Proteinase K (Roche Cat. No. 03 115 828 001)
4. Phenol/chloroform/IAA (Sigma-Fluka Cat. No. 777617; 100 ml)
5. Chloroform/IAA (Sigma-Fluka Cat. No. 25666; 100 ml)
6. Sodium acetate – AnalaR or similar high quality
7. Isopropanol or ethanol
8. Eppendorf 1.5 ml Phase-lock gels, light (VWR International Cat No. 713–253) or
9. Eppendorf 15 ml Phase-lock gels, light (VWR International Cat. No. 713–2537)

3 Methods

3.1 Initial Phage Infections

The infections carried out to obtain semi-confluent phage lysates is carried out as per usual. The phage particles are first obtained from the top agarose layer by addition of lambda diluent and scraped off into a falcon tube. For every 3 ml of lambda diluent added, 200 µl of chloroform is added. The resulting lyaste is shaken on a whirlimixer to help free the phage from the lysed

bacteria. This is followed by methods to burst the phage open so as to release the phage DNA and purify away from the degraded genomic DNA fragment contamination. The two purification methods commonly used in our laboratory are now described in more detail below. A brief description of preparing the high titre plate lysates is included before hand for the sake of completeness.

3.2 Obtaining High Titre Phage Stocks from 1 or 2 L-Agar/Agarose Plates

1. Bacterial strain required for phage infection is put up overnight in 5 ml of L-broth (+ any supplements if required) and shaken at 37 °C. L-agar plates containing 1.4% agar are prepared (with any supplements and antibiotics added).
2. Next day, one bottle of SOFT L-agarose (0.35 to 0.7% maximum-lower values may be better) is melted and cooled at 56 °C ready for use later. Five minutes before use (in step 5 below), cool to 42 °C.
3. Serial dilutions of the phage stock are made in lambda buffer from −1 to −5 (depending upon titre if known). Ten microlitres of aliquots are added to 15 ml falcon tubes.
4. Add 200 µl of O/N culture to each tube and mixed gently. The tubes are incubated at 37 °C for 20 min to allow phage absorption onto host bacterium.
5. Add 3 ml of 0.5% L-agar to the phage/bacteria mix. It is poured onto a 1.4% L-agar plate immediately. Plates are incubated until zones of phage confluence are seen.
6. Add 3 ml lambda buffer to the plate and leave O/N at 4 °C.
7. Next day, scrape off top agar plus the buffer. Add to a 50 ml falcon tube and add 50 µl of chloroform. Shake for 1 min. Spin down supernatant at 4,000 rpm for 30 min in a bench centrifuge.
8. Collect supernatant and filter sterilise to remove any insoluble matter (a 45 or 70 µM filter is preferable). It is now ready for the phage DNA purification protocols.

3.3 Isolation of Purified Bacteriophage DNA from Phage Particles

Three alternative methods to prepare the phage DNA are described. Two are based on a Promega kit-based method, but differ only in the use of a Promega Vacuum Manifold, while the alternative uses a syringe barrel and plunger. The third method uses phase-locked gels in combination with phenol–chloroform extractions. The latter gives slightly better yield so may be the method of choice for some phage preparations where the phage *burst size is smaller*.

3.3.1 METHOD ONE—Using the Promega Wizard Lambda Preps DNA Purification System (Product Number: A7290)

This method uses a 3–5 ml Luer-Lok syringe. If a Promega Vac-Man Laboratory Manifold or similar is available then follow the modified method described below (modified version ONE), as this variation will give slightly better consistent results as well as allow simultaneous purification of many samples.

1. Add 40 µl of nuclease mixture to 5 ml of phage supernatant (I use typically 5 ml of supernatant from just one plate. If volume is less than 5 ml then top up with lambda diluent). Incubate 37 °C for 15 min.
2. Add 4 ml of phage precipitant, mix gently and leave for 30 min on ice.
3. Spin 10, 000 × g for 10 min.
4. Carefully remove the supernatant and discard it. Resuspend the spun down pellet in 500 µl of phage buffer. To remove any DNAase activity left over from the nuclease step (which is present in the nuclease mixture to degrade contaminating genomic DNA and RNA), add proteinase K to a final concentration of 0.5 mg/ml and incubate at 37 °C for 10 min.
5. Transfer the resuspended phage to an Eppendorf tube and spin for 10 s so as to remove any insoluble particles. Transfer supernatant to a fresh Eppendorf and add 1 ml of Promega purification resin (shake before use). The sample is now ready for use with a 5 ml syringe so as to continue the purification. Attach a Promega mini-column to the syringe.
6. Pipette the resin/lysate into the syringe barrel. Insert the plunger slowly and gently push the solution into the mini-column. Detach the syringe from the mini-column and in a further syringe add 2 ml of 80% isopropanol. Pass this through the mini-column as well by pushing down on the plunger.
7. Transfer the mini-column to a 1.5 ml Eppendorf. Centrifuge the mini-column for 30 s to dry the resin.
8. Transfer column to a fresh Eppendorf. Elute purified phage DNA with 100 µl of TE buffer (or E buffer which is 10 mM Tris [pH 8.5]) that had previously been heated to 80 °C. Immediately spin the mini-column in the centrifuge at 14,000 rpm for 1 min and store the collected bacteriophage DNA at 4 °C.

If a Promega Vac-Man Laboratory Manifold or similar is available then a simple modification to the procedure can be carried out.

3.3.2 Modified Version of Method ONE Using a Promega Vac-Man Laboratory Manifold

1. Add 40 µl of nuclease mixture to 5 ml of phage supernatant (I use typically 5 ml of supernatant from just one plate. If volume is less than 5 ml then top up with lambda diluent). Incubate 37 °C for 15 min.
2. Add 4 ml of phage precipitant, mix gently and place on ice for 30 min.
3. Spin 10, 000 × g for 10 min.
4. Carefully decant the supernatant and resuspend the pellet in 500 µl of phage buffer. To remove any DNase I activity left over from the nuclease step (which was present in the nuclease mixture to degrade contaminating genomic DNA

and |RNA), add proteinase K to a final concentration of 0.5 mg/ml and incubate at 37 °C for 10 min.

5. Transfer the resuspended phage to an Eppendorf tube and spin for 20 s so as to remove any insoluble particles. Transfer supernatant to a fresh Eppendorf and add 1 ml of Promega purification resin (shake before use).
6. For each phage lysate prep, attach one supplied syringe barrel to the Luer-Lok extension of the Promega Wizard mini-column. This entire assembly is then simply attached to the Promega Vacuum Manifold by the tip at the end of the mini-column.
7. Pipette the resin/lysate mix into the syringe barrel and apply the vacuum to draw this suspension into the mini-column. Once all the contents have entered the column, switch off the vacuum line.
8. Add 2 ml of 80% isopropanol to the syringe barrel and apply vacuum once again. This step washes the column.
9. Dry the resin by applying the vacuum for 30 s exactly after all the isopropanol has entered the mini-column. It is important not to over-dry the column at this stage.
10. Transfer the mini-column to a new Eppendorf. The column has been designed to fit on top of a 1.5 or 2.0 ml Eppendorf. Elute-purified phage DNA with 100 µl of TE buffer (or E buffer which is 10 mM Tris [pH 8.5]) that had previously been heated to 80 °C. Immediately spin the mini-column in the centrifuge at 14,000 rpm for 1 min and store the collected bacteriophage DNA at 4 °C.

3.3.3 Method Two—Using the Phase Lock Gel System (Phenol–Chloroform Method)

These columns are available from VWR and are supplied in two sizes, but the method below describes using the smaller 1.5 ml Eppendorfs.

1. Transfer 1.8 ml aliquots (assume 2.5–3.0 ml of phage lysate per plate of confluent phage) of phage lysate to a 15 ml falcon tube. Add 18 µl of 1 mg/ml DNase I and 8 µl of 12.5 mg/ml RNase A. Mix and incubate at 37 °C for 30 min. This removes genomic contaminants from the phage lysate.
2. Add 46 µl of 20% SDS and 18 µl of 10 mg/ml proteinase K to the sample, mix and incubate for a further 30 min at 37 °C. Aliquot 500 µl volumes into 4 × 1.5 ml Phase-lock gel Eppendorfs.
3. Extract aliquoted samples with 0.5 ml of phenol:chloroform: isoamyl alcohol (25:24:1). Spin 5 min at 1,500 × g to separate the phases.
4. Remove top aqueous phase into a fresh Phase-lock gel tube and *repeat* the above step.
5. Remove aqueous phase into a fresh Phase-lock gel tube and extract once with chloroform:IAA (24:1). Centrifuge for 5 min at 6,000 × g.

6. Transfer the aqueous phase to a 1.5 ml Eppendorf tube and add 45 µl of 3 M sodium acetate (pH 5.2) and 500 µl of 100% isopropanol (can use two volumes of ethanol as an alternative instead). Leave DNA to precipitate at room temperature for 20 min.
7. Spin at 14,000 rpm for 20 min and wash DNA pellet twice with 70% ethanol prior to drying.
8. Resuspend DNA to a final total volume of 200 µl (50 µl per Eppendorf) with TE or E buffer and transfer DNA to a sterile Eppendorf. Store at 4 °C.

3.3.3.1 Purifying Phage DNA from a Larger Volume of Lysate and Utilising the Larger 15 ml Phase-Lock Gel Falcon Tubes

The method detailed above uses the 1.5 ml Phase-lock gel Eppendorfs, but if a lysate volume of 3.6 ml is treated to generate a larger volume of phage DNA, then 15 ml tubes can be used instead and all volumes are doubled in the first three steps (i.e. 36 µl of 1 mg/ml DNase I; 16 µl of RNase A; 92 µl of 20% SDS and 36 µl of proteinase K). For this volume in one 15 ml Phase-lock gel

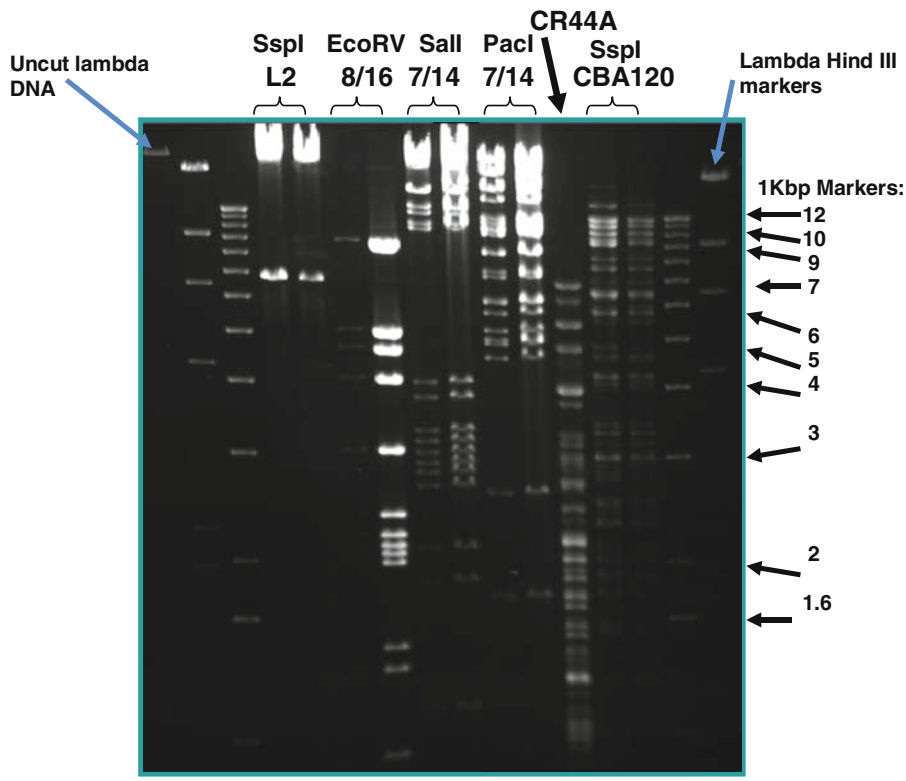


Fig. 1.1. A range of bacteriophage DNA was prepared by the Phase-lock gel procedure. Starting volume was either 0.8 or 3.2 ml. L2, 8/16 and 7/14 are lytic phage DNA preps obtained from *Pseudomonas aeruginosa* and cut with a variety of restriction enzymes. CR44A was a lytic phage obtained from infection of *Citrobacter rodentium*, while CBA120 is a lytic phage of *Escherichia coli*. For a number of samples, DNA had been prepared on more than one occasion and using different volumes of starting material (i.e. phage 8/16). Phage kindly provided by Betty Kutter, Seamus Flynn and Ana Luisa Toribio.

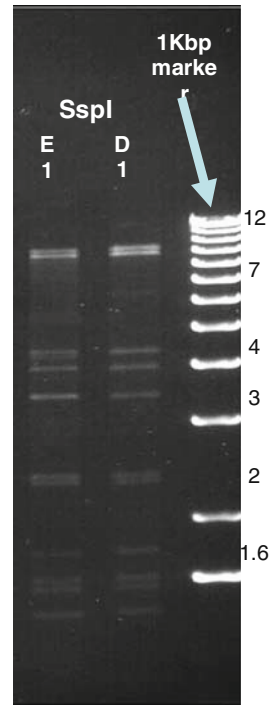


Fig. 1.2. Bacteriophage DNA from *Salmonella typhi* lytic phage E1 and D1 prepared using the Promega lambda phage kit. Yield was found to be on an average of about 10% lower than that obtained using the Phase-lock gel system. DNA was digested using SspI. Phage kindly provided by Colindale typing laboratories.

tube, we would need to use 3.8 ml of phenol/chloroform/IAA, etc. and scale up all subsequent steps and adjust volumes when required. For example, after the chloroform:IAA step make up aqueous volume to 4 ml with E buffer, add 360 μl of 3 M sodium acetate and 4 ml of 100% isopropanol. Incubate to precipitate DNA as before.

4 Concluding Comments

All these methods should yield DNA of at least 99% purity such that they can be used for restriction digestions and sequencing. Yields of DNA can be very good and approach 10 μg per 1.8 ml of cleared lysate starting material. Typical results with a variety of phage are shown in Figs. 1.1 and 1.2. Though the gel shows DNA prepared from lytic page, the procedures mentioned would be suitable for lysogenic phage grown in suitable hosts and conditions to elicit the lytic cycle (i.e. lambda gt11, lambda phage, stx I or II, etc.).

Chapter 2

Approaches to the Compositional Analysis of DNA

Richard A. Manderville and Andrew M. Kropinski

Abstract

DNA base compositional analysis is something which is rarely undertaken today, but it is still a useful criterion for phage taxonomy. A variety of techniques are described including hydrolysis of the DNA to the level of bases or nucleosides and separation by paper chromatography or HPLC. Spectroscopic and spectrofluorometric procedures are also outlined.

Key words: Base composition, G+C, unusual bases, buoyant density, melting temperature, spectra ratios, UV spectroscopy, spectrofluorometry, HPLC, phosphodiesterase, deoxyribonucleosides, nuclease P1.

1 Introduction

The base composition of DNA, generally referred to as the mol% guanine plus cytosine or %GC, is one of the defining characteristics of viral DNAs as outlined by the International Committee on Taxonomy of Viruses *(1)*. A number of procedures are available for determining the base composition of nucleic acids: these include hydrolysis, separation and quantitation of individual bases, nucleosides or nucleotides; spectrophotometric and spectrofluorometric procedures; and approaches based on the buoyant density of DNA in solutions of cesium salts. Some of these procedures require specialized expensive equipment, but in this chapter we also describe inexpensive alternative protocols.

1.1 Classical Approach

The classical approach pioneered by Gerry Wyatt *(2)* and described in depth by Aaron Bendich *(3)* involves hydrolysis of milligram quantities of DNA in concentrated perchloric (100 °C/60 min) or formic acids (175 °C/30 min) followed by

Martha R. J. Clokie, Andrew M. Kropinski (eds.), *Bacteriophages: Methods and Protocols, Volume 2: Molecular and Applied Aspects, vol. 502*, © 2009 Humana Press, a part of Springer Science+Business Media
DOI 10.1007/978-1-60327-565-1_2 Springerprotocols.com

the separation of the bases by descending paper chromatography using freshly prepared 2-propanol–HCl–water (65:17:18, vol/vol). The UV-absorbing spots are excised and the bases eluted from the paper in 0.1 M of HCl. The concentration of each can be easily measured spectrophotometrically based on their specific extinction coefficients (*see* **Notes 1, 2**, and **3**).

1.2 Spectrophotometric Approaches

The advantages of the following three procedures are that they are theoretically nondestructive and require small amounts of DNA (20–40 μg/ml DNA). The first technique only requires a UV spectrophotometer.

1.3 Spectral Properties of DNA at pH 3—Two Approaches

The spectrophotometric properties of DNA at pH 3 were first exploited by Fredericq et al. *(4)*, who noted a log–log relationship between the 260:280 spectral ratio, in 0.05 M of acetic acid, and the mol%A + T of DNA. This was subsequently modified by Maiti and Nandi *(5)*, who determined the circular dichroic properties of DNA in a citrate–phosphate buffer and noted a linear relationship between the %GC and the ellipticity ratio ($\theta_{260}/\theta_{280}$):

$$\text{Mol\%G} + \text{C} = \frac{\theta_{260}/\theta_{280} + 1.54}{0.018}$$

(*see* **Notes 4, 5**, and **6**)

1.4 Melting Temperature

When DNA is heated, the chromaticity at 260 nm shows a sharp increase during strand separation. The midpoint of the absorbance increase, referred to as the melting temperature (T_m) is dependent on the salt concentration and the %GC *(6, 7)*. In most cases, DNA is dialyzed against "standard saline citrate" (1X SSC; 0.15 M of NaCl–0.015 M of Na_3citrate [pH 7.0]) prior to use. In this "buffer," the %GC can be calculated using the following equations:

$$\%\text{GC} = 1.99(T_m - 66.0) \qquad (8)$$

$$\%\text{GC} = 2.44(T_m - 69.4) \qquad (9)$$

(*see* **Notes 7, 8, 9, 10**, and **11**)

The use of melting temperatures and buoyant density determinations (see below) were the major techniques employed to determine the %GC from the 1970s into the next century. The techniques described in Methods in Enzymology employ a Gilford Spectrophotometer which is no longer manufactured. The products of five spectrophotometer manufacturers (Beckman Coulter, Inc., Fullerton, CA, http://www.beckmancoulter.com/; Cecil BioAquarius; Pocklington, York, United Kingdom, http://www.wolflabs.co.uk/; PerkinElmer, Inc., Boston, MA, http://www.perkinelmer.com/;

Simadzu, Columbia, MD, http://www.ssi.shimadzu.com/; and Varian [Cary], Walnut Creek, CA, http://www.varianinc.com) are all equipped easy for analysis of the DNA T_m. These include microcuvettes, Peltier temperature controllers and often software.

1.5 Spectrofluorometric Techniques

A number of fluorescent dyes exhibit differential binding to DNAs of differing GC-contents and this likewise has been exploited to determine the molar percentage of these bases in nucleic acids. Daxhelet et al. *(10)* noted that olivomycin (Zyf Pharm Chemical, China) showed a strong linear relationship between GC-contents and fluorescence. The presence of the modified bases 5-hydroxymethylcytosine and 5-hydroxymethyluracil results in deviations from linearity. As with the above-mentioned techniques, one must construct standard curves using DNAs of known base composition.

1.6 Buoyant Density

The buoyant density of DNA in solutions of cesium salts is also affected by its %GC content *(11,12)*, a fact that has been exploited to determine the guanine and cytosine content of DNA. As with the T_m determinations, this procedure requires access to expensive equipment: in this case a Beckman analytical ultracentrifuge (models Optima™ XL-1, Optima XL-A or ProteomeLab XL-A) with UV optics.

The %GC can be calculated using the following equations:

$$\%GC = \frac{100(\rho_{CsCl} - 1.660)}{0.098} \qquad (11, 13)$$

$$\%GC = 1038.47(\rho_{CsCl} - 1.6616) \qquad (9)$$

(*see* **Note 12**)

1.7 HPLC

Base composition analysis using HPLC with UV–Vis detection requires enzymatic digestion of the DNA prior to HPLC analysis of the monodeoxynucleosides *(14,15)*. Enzymatic digestion of DNA depends on the enzyme activity, the amount of DNA used for the enzymatic digestion, DNA concentration, and the incubation time. Incomplete enzymatic digestion probably contributes to interlaboratory variations in compositional analysis of DNA by HPLC. In a relatively recent publication by Huang et al. *(16)*, three enzymatic DNA digestion protocols were examined with the goal of maximizing release of normal nucleosides per microgram of DNA and to minimize oxidation of dG and DNA during sample preparation and handling. The optimal DNA digestion protocol is described below.

2 Materials

1. Purified bacteriophage DNA in TE buffer (*see* **Note 13**).
2. 1 mM $CaCl_2$, 10 mM Mg^{2+}, 10 mM Tris (pH 8.5), 40 μM diethylenetriamine pentaacetic acid (DTPA, Sigma-Aldrich).
3. 3 M sodium acetate (pH 5.2).
4. DNase I *Escherichia coli* alkaline phosphatase and, snake venom phosphodiesterase I of the highest quality (these have enzymes for DNA digestion can be purchased from Roche Molecular Bio-Chemicals (Indianapolis, IN, USA)).
5. 1 M Tris–HCl (pH 8.0).
6. Block heater set at 37 °C.
7. 0.1 M triethylammonium acetate (pH 6.5) containing 5% CH_3CN (buffer A) and 0.1 M triethylammonium acetate (pH 6.5) containing 65% CH_3CN (buffer B).
8. Ultrafree-MC membrane (nominal molecular weight limit 5,000; Millipore Corp., Billerica, MA; http://www.millipore. com/).
9. Reversed phase HPLC on analytical size C18 columns (e.g., 5 μm Agilent ZORBAX Eclipse XDB C18 column, 4. 6 mm × 150 mm; Agilent Technologies/Quantum Analytics, Inc., Foster City, CA, http://www.chem.agilent.com/).
10. HPLC equipment.

3 Methods

1. Thirty micrograms of DNA is digested in 100 μl solution containing 1 mM $CaCl_2$, 10 mM Mg^{2+}, 10 mM Tris (pH 8.5), 40 μM diethylenetriamine pentaacetic acid (DTPA, Sigma-Aldrich) and 40 units DNase I for 1 h at 37 °C.
2. Add 1 μl of alkaline phosphatase (AP, 1 unit/μl) followed by one hour's incubation.
3. Add 1 μl of phosphodiesterase I (PDE, 0.01 unit/μl each) to the reaction mixture and incubate for an additional hour to ensure the completeness of DNA digestion.
4. After digestion, reaction mixtures should be filtered to remove enzymes before injection onto the HPLC column (*see* **Note 14**).
5. The deoxynucleosides produced can be analyzed by reversed phase HPLC on analytical size C18 columns (e.g., 5 μm Agilent ZORBAX Eclipse XDB C18 column, 4. 6 mm × 150 mm; Agilent Technologies/Quantum Analytics, Inc., Foster City,

CA, http://www.chem.agilent.com/) with 0.1 M triethylammonium acetate (pH 6.5) containing 5% CH_3CN (buffer A) and 0.1 M triethylammonium acetate (pH 6.5) containing 65% CH_3CN (buffer B) operated at a flow rate of 1 ml/min and ambient temperature *(16)*. A common system consists of isocratic elution with 95% buffer A and 5% buffer B. While under these conditions, the order of elution is dC, dG, T, dA, authentic standards of the four deoxynucleosides should be used for comparison and to generate standard curves that can be used for quantification purposes (*see* **Note 15**).

6. The amounts of each deoxynucleoside can be determined by integration to give the areas under each peak in the HPLC trace. These areas must be divided by the following extinction coefficients at 254 nm to take into account the different absorptions of the deoxynucleosides at the detection wavelength *(17)*. Levels of the normal nucleosides are quantified using the standard curves derived from standards. The extinction coefficients at 254 nm: dC (6×10^3), dG (13.5×10^3), T (7×10^3), and dA (14.3×10^3).

4 Notes

1. Some modified bases such as *N*-putrescinylthymine and hydroxymethylcytosine are destroyed during $HClO_4$ hydrolysis.
2. Great care should be taken when carrying out hydrolyses in formic acid since extreme pressure builds up within the vials and they can spontaneously shatter. After hydrolysis it is recommended that the liquid in the vials be frozen prior to opening.
3. This technique could be miniaturized if thin layer chromatography (TLC) on cellulose layers on plastic or glass plates was employed. Do not use prepared plates which contain UV-absorbing materials. This approach would still require a minimum of 200 μg of DNA per experimental determination.
4. When employing the procedure of Fredericq et al., we would recommend the use of Maiti and Nandi's buffer system: 20 mM citric acid–10 mM disodium phosphate (pH 3.0).
5. In both of these procedures, it is strongly recommended that control DNA samples be employed to standard curves of A_{260nm}/A_{280nm} (or $\theta_{260}/\theta_{280}$) versus %GC. Sigma-Aldrich (St Louis, MO, http://www.sigmaaldrich.com) sells purified double-stranded DNAs from ϕX174 (44.76 %GC), M13mp18 (42.36%), coliphage lambda (49.86%), *Clostridium perfringens* (28.57%), *Micrococcus luteus* (65.1% based

on sequences in GenBank and 72.4% based on buoyant density), and *Escherichia coli* strain B (50.8%). Genomic DNA is also available from the American Type Culture Collection from a wide variety of bacterial strains (http://www.atcc.org/common/products/PurifiedDNA.cfm). The latter is very expensive and only available in 10 μg aliquots. As an alternative, it is recommended that you purify DNA from a bacterial strain which has been sequenced (see NCBI, http://www.ncbi.nlm.nih.gov/ genomes/lproks.cgi).

6. It is not known how modified bases, and DNA linearity/circularity will affect the spectral ratios.
7. In determining the T_m, it is always advisable to include a DNA sample of known %GC as a control.
8. Certain modified bases can influence the T_m resulting in a higher (5-methylcytosine, 2-aminoadenine, and *N*-putrescinylthymine) or lower (α-glutamylthymine, 5-hydroxypentyluracil, uracil, and hydroxymethyluracil) than expected value *(18)*.
9. For high GC-content DNAs, it is recommended that it is dialyzed against 0.1X SSC prior to use.
10. The following equation holds between 0.05 and 3X SSC for the relationship between the T_m and the %GC (N.B. for every 10-fold increase or decrease in SSC concentration ([SSC]) the T_m changes by 16.3 °C:

$$\%GC = (T_m - 16.3 \log_{10} \text{Relative [SSC]}/50.2) - 0.990 \quad (8)$$

where Relative [SSC] = Concentration of SSC relative to 0.1X SSC

11. In 0.0025 M of Na_2HPO_4–0.005 M of NaH_2PO_4–0.001 M of EDTA (pH 6.8), the T_m is 20.0 °C lower than in SSC *(19)*
12. As with the T_m determinations, the presence of unusual or modified bases can contribute to a discrepancy between the theoretical and observed %GC values. The presence of *N*-putrescinylthymine or 5-methylcytosine decreases; and, uracil, 5-dihydroxypentyluracil, 5-hydroxymethyluracil, 2-aminoadenine, and α-glutaminylthymine increase the buoyant densities of DNAs *(18)*.
13. The DNA should be free of RNA contamination. While the 260:280 ratio may give an indication of RNA contamination, agarose gel electrophoresis will give you a better idea about the quality of the DNA.
14. Digested DNA may be filtered through an Ultrafree-MC membrane (nominal molecular weight limit 5,000; Millipore Corp., Billerica, MA, http://www.millipore.com/) by centrifugation (9,000 rpm).

15. Modification of the HPLC protocol may be required to resolve modified deoxynucleosides, such as methylated analogs, from their normal compounds.

References

1. van Regenmortel, M.H.V., C.M. Fauquet, D.H.L. Bishop, E.B. Carstens, M.K. Estes, S.M. Lemon, J. Maniloff, D.J. McGeoch, et al. 2000. Virus Taxonomy: Classification and Nomenclature of Viruses – Seventh Report of the International Committee on the Taxonomy of Viruses. Academic Press, New York.
2. Wyatt, G.R. 1951. The purine and pyrimidine composition of deoxypentose nucleic acids. Journal of Biochemistry *48*:584–590.
3. Bendich, A. 1957. Methods for characterization of nucleic acids by base composition. Methods in Enzymology *3*:715–723.
4. Fredericq, E., A. Oth, and F. Fontaine. 1961. The ultraviolet spectrum of deoxyribonucleic acids and their constituents. Journal of Molecular Biology *3*:11–17.
5. Maiti, M. and R. Nandi. 1987. Spectropolarimetric determination of the guanine-cytosine content of DNA. Analytical Biochemistry *164*:68–71.
6. Marmur, J. and P. Doty. 1962. Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature. Journal of Molecular Biology *5*: 109–118.
7. Schildkraut, C.L. and S. Lifson. 1965. Dependence of the melting temperature of DNA on salt concentration. Biopolymers *3*:195–208.
8. Mandel, M., L. Igambi, J. Bergendahl, M.L. Dodson, Jr., and E. Scheltgen. 1970. Correlation of melting temperature and cesium chloride buoyant density of bacterial deoxyribonucleic acid. Journal of Bacteriology *101*: 333–338.
9. De Ley J. 1970. Reexamination of the association between melting point, buoyant density, and chemical base composition of deoxyribonucleic acid. Journal of Bacteriology *101*:738–754.
10. Daxhelet, G.A., M.M. Coene, P.P. Hoet, and C.G. Cocito. 1989. Spectrofluorometry of dyes with DNAs of different base composition and conformation. Analytical Biochemistry *179*:401–403.
11. Schildkraut, C.L., J. Marmur, and P. Doty. 1962. Determination of the base composition of deoxyribonucleic acid from its buoyant density in CsCl. Journal of Molecular Biology *4*:430–443.
12. Sueoka, N., J. Marmur, and P. Doty. 1959. Dependence of the density of deoxyribonucleic acids on guanine-cytosine content. Nature *183*:1429–1431.
13. Mandel, M., C.L. Schildkraut, and J. Marmur. 1968. Use of CsCl density gradient analysis for determining the guanine plus cytosine content of D. Methods in Enzymology *12*: 184–195.
14. Kuo, K.C., R.A. McCune, C.W. Gehrke, R. Midgett, and M. Ehrlich. 1980. Quantitative reversed-phase high performance liquid chromatographic determination of major and modified deoxyribonucleosides in DNA. Nucleic Acids Research *8*:4763–4776.
15. Wakizaka, A., K. Kurosaka, and E. Okuhara. 1979. Rapid separation of DNA constituents, bases, nucleosides and nucleotides, under the same chromatographic conditions using high-performance liquid chromatography with a reversed-phase column. Journal of Chromatography *162*:319–326.
16. Huang, X., J. Powell, L.A. Mooney, C. Li, and K. Frenkel. 2001. Importance of complete DNA digestion in minimizing variability of 8-oxo-dG analyses. Free Radical Biology & Medicine *31*:1341–1351.
17. Connolly, B.A. 1991. Oligonucleotides containing modified bases., *In* F. Eckstein (Ed.), Oligonucleotides and Analogues A Practical Approach. Oxford University Press, New York.
18. Warren, R.A.J. 1980. Modified bases in bacteriophage DNA. Annual Review of Microbiology *34*:137–158.
19. Mandel, M. and J. Marmur. 1968. Use of ultraviolet absorbance-temperature profile for determining the guanine plus cytosine content of DNA. Methods in Enzymology *12*: 195–206.

Chapter 3

Determination of Bacteriophage Genome Size by Pulsed-Field Gel Electrophoresis

Erika Lingohr, Shelley Frost and Roger P. Johnson

Abstract

Standard agarose gel electrophoresis is extensively used to resolve DNA fragments from 0.2 to 40–50 kb. Larger fragments of genomic DNA or whole viral genomes can only effectively be resolved by pulsed-field gel electrophoresis (PFGE), which extends the range of molecular separation from 200 bp to 12 Mb.

Key words: Pulsed-field gel electrophoresis, PFGE, agarose, genome sizing, RFLP.

1 Introduction

Standard agarose gel electrophoresis is extensively used to resolve DNA fragments from 0.2 to 40–50 kb. Larger fragments of genomic DNA or whole viral genomic DNAs co-migrate as a result of reptation (the snake-like motion of entangled polymers). Pulsed-field gel electrophoresis (PFGE) extends the range of molecular separation from 200 bp to 12 Mb, thus it is the technique of choice for analyzing prokaryote-sized genomes. Indeed, it is the basis of PulseNet—the molecular screening of pathogens on the basis of restriction endonuclease cleavage patterns *(1, 2)*. With respect to viruses, PFGE has been used to characterize large viral genomes *(3, 4)*, to analyze the integration of prophages *(5)*, to study the replicative formation of concatemers, and to distinguish between circular and linear genomes [for review, *see (6)*]. Additionally, in the area of viral metagenomics, PFGE has been employed to analyze viral diversity in marine *(7)* and hypersaline environments *(8)*, and in human fecal matter *(9)* [for review, *see (10)*].

Martha R. J. Clokie, Andrew M. Kropinski (eds.), *Bacteriophages: Methods and Protocols, Volume 2: Molecular and Applied Aspects, vol. 502*, © 2009 Humana Press, a part of Springer Science+Business Media
DOI 10.1007/978-1-60327-565-1_3 Springerprotocols.com

The following procedure is a modification of a PFGE protocol for analysis of DNA from bacterial cells *(1)* that has been optimized for purified large coliphages. The phages are prepared in an agarose plug, and are lysed within the plug by treatment with detergent and a proteinase. The plug is then washed and a slice of it is loaded into an agarose gel for PFGE. The pulsed electrical field allows entry and mobility of fragments of DNA larger than can normally migrate in standard agarose electrophoresis (*see* **Fig. 3.1**). If required, whole genomic DNA can be digested with one or more restriction endonucleases prior to PFGE, to generate smaller fragments for restriction fragment length polymorphism (RFLP) analysis. It is often more suitable to use intact, purified phages for genome size estimation by PFGE analysis than extracted phage DNA, because there is less shearing of genomic

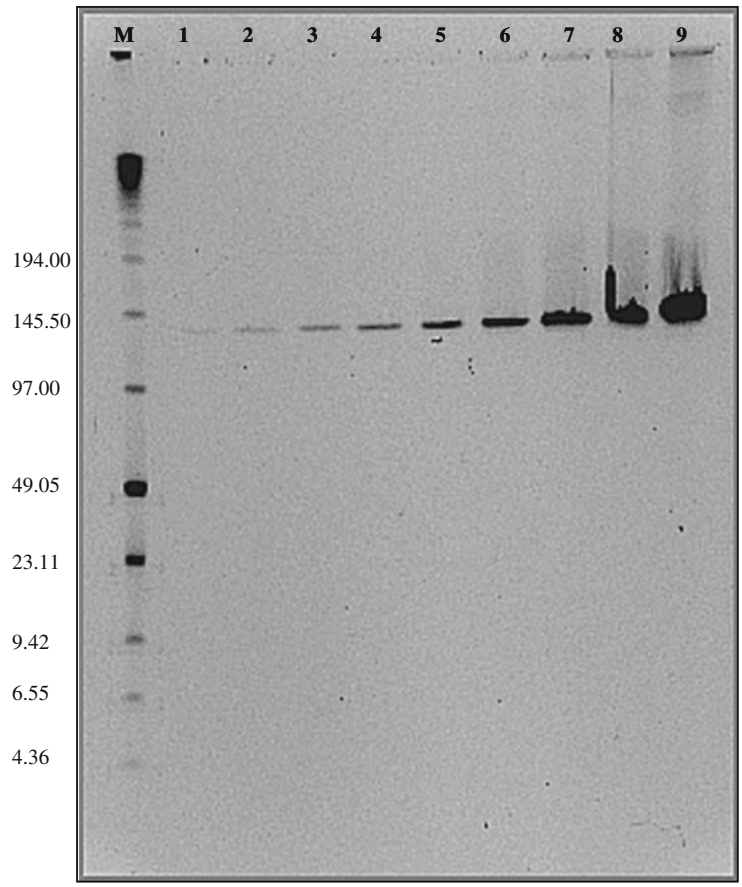


Fig. 3.1. Pulsed field gel electrophoresis of coliphage rV5, purified by cesium chloride density gradient ultracentrifugation, and prepared in agarose plugs without DNA extraction, as described in the text. Lane M, low range PFGE marker DNA. Lanes 1–9 were loaded with slices from plugs containing twofold dilutions of the phages, with the highest number $(1.38 \times 10^8 \text{ pfu})$ in the slice loaded in Lane 9.

nucleic acid. Phages which have been purified by density gradient ultracentrifugation need to be dialyzed against a low salt buffer to remove cesium chloride or sucrose before processing *(11)*. Although this protocol has been optimized for sizing phage genomes of approximately 100–150 kb, it is adaptable to smaller and larger genomes, with the adjustment of the number of phages prepared in the agarose plug.

2 Materials

2.1 Equipment

1. Refrigerator set at 4 °C.
2. Incubator set at 50 °C.
3. Shaking water bath set at 54 °C.
4. Microcentrifuge tubes, sterile.
5. Small flat spatula, sterile.
6. Scalpel blade or gel knife, sterile.
7. Screw-capped centrifuge tubes, 15 ml, sterile.
8. PFGE system (BioRad CHEF-DRIII or equivalent) with plug casting molds.
9. Pipettors and tips for volumes between 20 µl and 10 ml, sterile.
10. Analytical balance.
11. Hot plate or microwave.
12. Ultraviolet illuminator and a gel documentation system.

2.2 Reagents and Buffers

Molecular biology grade reagents, buffers, and water from commercial suppliers with good quality control are recommended to reduce run-to-run variations. Evaluation of reagents with the same specifications, but from different suppliers, is recommended for the same reason. Many of the following formulations are readily available from commercial sources.

1. Purified phages at suitable titer, 10^6–10^9 pfu/ml, typically 10^7 pfu/ml (**Note 1**).
2. Molecular biology grade water, sterile, de-ionized, and free of nucleases and DNA and RNA.
3. 1.0 M Tris (pH 8.0).

Per 500 ml:	Tris [Tris-hydroxymethyl aminomethane]	60.57 g
	Water to	350 ml
	Adjust pH to 8.0	
	Add water to	500 ml

4. 0.5 M EDTA (pH 8.0)

Per 500 ml:	EDTA (disodium ethylenediaminetetraacetate)	93.05 g
	Water to	350 ml
	Adjust pH to 8.0	
	Add water to	500 ml

5. Tris-EDTA (TE) buffer, 1X, (10 mM Tris and 1 mM EDTA [pH 8.0]).
   Per 100 ml: 1 M Tris, pH 8.0 1 ml
   0.5 M EDTA, pH 8 0.2 ml
   Water to 100 ml
6. Tris-borate-EDTA (TBE), buffer, 5X (0.45 M Tris borate and 0.01 M EDTA).
   Per 1,000 ml: Tris base 54 g
   Boric acid 27.5 g
   0.5 M EDTA, pH 8.0 20 ml
   Add water to 1 L
7. Tris-borate-EDTA (TBE), buffer, 0.5X (45 mM Tris borate and 1 mM EDTA).
   Per 1,000 ml: 5X TBE 200 ml
   Water to 1 L
8. Phage suspension (PS) buffer (0.1 M Tris and 0.1 M EDTA [pH 8.0]).
   Per 100 ml: 1.0 M Tris, pH 8.0 10 ml
   0.5 M EDTA, pH 8.0 20 ml
   Water to 100 ml
9. Plug agarose (1.2% SeaKem Gold Agarose [Cambrex Corp.; http://www.cambrex.com/default.asp], 1X TE Buffer).
   Per 100 ml: SeaKem Gold Agarose 1.2 g
   1X TE Buffer 100 ml
   Heat until dissolved then hold at 50 °C
10. Phage lysis (PL) buffer (50 mM Tris, 50 mM EDTA, and 1% (w/v) SDS).
    Per 100 ml: 1.0 M Tris, pH 8.0 5 ml
    0.5 M EDTA, pH 8.0 10 ml
    SDS 1 g
    Water 85 ml
11. Proteinase K solution, 20 mg/ml.
    Per 1 ml: Proteinase K 20 mg
    Sterile nuclease-free water 1 ml
12. Ethanol, 70% (v/v).
13. PFGE agarose (1% SeaKem Gold Agarose, 0.5X TBE).
    Per 1,000 ml: SeaKem Gold Agarose 1.2 g
    0.5X TBE Buffer 120 ml
    Water to 1 L
    Heat until dissolved and cool to 50 °C before casting gel.
14. PFGE low range DNA Marker in agarose plugs (New England Biolabs; Ipswich, MA, http://www.neb.com/; Catalogue No. N0350S), approximately 0.13–194 kb (**Note 2**).
15. Ethidium bromide solution, 1X, 0. 5–1 μg/ml
    Per 1,000 ml: Ethidium bromide 0.5 to 1.0 mg
    Distilled Water 1,000 ml

3 Methods

3.1 Plug Preparation

1. If necessary, dialyze the purified phage preparation against three changes of PS buffer, to remove cesium chloride or sucrose (**Note 1**).
2. Assemble the plug casting molds, and label the wells.
3. Prepare the plug agarose in a volume of 0.5–1 ml per phage preparation, and hold at 50–54 °C in a heating block or water bath.
4. Transfer 400 μl of dialyzed, purified phage preparation to labeled microcentrifuge tubes and warm the tubes in a heating block at 50 °C (**Note 3**).
5. Working with one phage preparation at a time, add 400 μl of molten plug agarose to the warmed phage preparation in the microcentrifuge tube, mix carefully by pipetting, ensuring no bubbles, and immediately transfer 250 μl of the mixture to fill a well of the plug casting mold.
   a. The remaining volume can be cast as additional plugs and stored in TE buffer, or allowed to solidify and stored at 4 °C.
6. Allow the plugs to solidify at 20–22 °C for 30 min or at 4 °C for 10–15 min.
7. Set up and label 15 ml screw-capped tubes for each plug sample.
8. Add 5 ml of PL buffer and 25 μl of proteinase K solution (20 mg/ml) to each tube.
9. Carefully open the wells of the casting mold and remove the plugs with a small, flat ethanol-sterilized spatula. Transfer each plug to the corresponding tube of PL buffer.
10. Place the tubes in a shaking water bath at 54 °C for 1.5–2.0 h. Ensure the water level of the bath is above that of the tubes.
11. For washing the plugs, heat sterile TE buffer to 54 °C in a water bath.
12. Remove the tubes containing the plugs from water bath, and carefully aspirate the buffer, ensuring that the plugs are saved.
13. Add at least 5 ml of warm, sterile TE buffer to each tube and incubate the tubes in a shaking water bath at 54 °C for at least 15 min.
14. Repeat steps 12 and 13 at least once, ensuring the plug is retained and the TE buffer is replaced.
15. At this stage, the plug can be stored at 4 °C until ready to load for PFGE.

a. *Optional*: At this stage, the plugs, or slices of the plug, can be treated with restriction endonucleases, if RFLP of the phage DNA is desired.

3.2 PFGE

1. Prepare 120 ml of 1% PFGE agarose in 0.5X TBE buffer.
2. Prepare 2.2 L of 0.5X TBE buffer, load it into the PFGE chamber and cool it to 14 °C.
3. Cool and cast the gel in the direction of the long axis using the appropriate well caster. Keep a small volume of the excess agarose at 50 °C in a heating block.
4. Allow the cast gel to solidify.
5. Remove the phage plugs from the TE buffer with a small, flat, ethanol-sterilized spatula, and with an ethanol-sterilized knife or scalpel blade, cut across the long axis of the plugs to make slices about one-fifth of the length of the plug. Store the remaining portions of the plugs at 4 °C in TE buffer.
6. Load the plug slices into the wells of the PFGE gel, ensuring that they touch the bottom and front walls of the wells. Load 2 mm slices of the marker DNA ladder and any other controls into respective wells in the gel.
7. Fill the wells with the remaining molten agarose to seal plugs in place, ensuring that there are no bubbles, and allow the gel to solidify.
8. Load the gel into the PFGE bed, containing the cooled buffer.
9. Run the gel at 6 V/cm for 18–20 h at 14 °C with incremental pulses of 2.2–54.2 s (**Note 4**).
10. Remove the gel, stain it in ethidium bromide solution (0.5 μg/mL) for 30 min, and then rinse it in de-ionized water until the unbound stain is cleared (30 min).
11. Examine the gel under ultraviolet light.
12. Capture, print, and save the image of the stained gel with a gel documentation system.
13. Estimate the size of the tested phage genomes visually by comparison with the DNA ladder and other controls (if included), or by analysis with appropriate software.

4 Notes

1. Phage can either be prepared from CsCl gradients (**Chapters 22 and Volume 2 Chapter 9**) or alternatively they can be taken from a fresh high titer supernatant from which bacterial debris is removed by centrifugation at 6,000 × g.
2. Other markers can be used. If required, they may need to be prepared as plugs before loading the PFGE gel.

3. Warming the suspension prevents immediate solidification of the added molten agar.
4. Pulse times may need to be vaired to optimize band separation.

References

1. Vivanco, A.B., J. Alvarez, I. Laconcha, N. Lopez-Molina, A. Rementeria, and J. Garaizar. 2004. Molecular genotyping methods and computerized analysis for the study of *Salmonella enterica*. Methods in Molecular Biology *268*:49–58.
2. Swaminathan, B., T.J. Barrett, S.B. Hunter, R.V. Tauxe, and T.F. CDC PulseNet. 2001. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. Emerging Infectious Diseases *7*:382–389.
3. Raoult, D., S. Audic, C. Robert, C. Abergel, P. Renesto, H. Ogata, B. La Scola, M.Suzan, and Claverie J.-M. 2004. The 1.2-Megabase Genome Sequence of Mimivirus. Science *Fundamentals of Measurement. 306*: 1344–1350.
4. Atterbury, R.J., P.L. Connerton, C.E. Dodd, C.E. Rees, and I.F. Connerton. 2003. Isolation and characterization of *Campylobacter* bacteriophages from retail poultry. Applied & Environmental Microbiology *69*: 4511–4518.
5. Iguchi, A., R. Osawa, J. Kawano, A. Shimizu, J. Terajima, and H. Watanabe. 2003. Effects of lysogeny of Shiga toxin 2-encoding bacteriophages on pulsed-field gel electrophoresis fragment pattern of *Escherichia coli* K-12. Current Microbiology *46*:224–227.
6. Serwer, P., S.J. Hayes, E.T. Moreno, D. Louie, R.H. Watson, and M. Son. 1993. Pulsed field agarose gel electrophoresis in the study of morphogenesis: packaging of double-stranded DNA in the capsids of bacteriophages. Electrophoresis *14*:271–277.
7. Fuhrman, J.A., J.F. Griffith, and M.S. Schwalbach. 2002. Prokaryotic and viral diversity patterns in marine plankton. Ecological Research *17*:183–194.
8. Diez, B., J. Anton, N. Guixa-Boixereu, C. Pedros-Alio, and F. Rodriguez-Valera. 2000. Pulsed-field gel electrophoresis analysis of virus assemblages present in a hypersaline environment. International Microbiology *3*: 159–164.
9. Breitbart, M., I. Hewson, B. Felts, J.M. Mahaffy, J. Nulton, P. Salamon, and F. Rohwer. 2003. Metagenomic analyses of an uncultured viral community from human feces. Journal of Bacteriology *185*:6220–6223.
10. Edwards, R.A. and F. Rohwer. 2005. Viral metagenomics. Nature Reviews Microbiology *3*:504–510.
11. Carlson, K. 2005. Working with bacteriophages: Common techniques and methodological approaches., *In* E. Kutter and A. Sulakvelidze (Eds.), Bacteriophages: Biology and Applications. CRC Press, Baco Raton, FL.

Chapter 4

Preparation of a Phage DNA Fragment Library for Whole Genome Shotgun Sequencing

Elizabeth J. Summer

Abstract

The most efficient method to determine the genomic sequence of a dsDNA phage is to use a whole genome shotgun approach (WGSA). Preparation of a library where each genomic fragment has an equal chance of being represented is critical to the success of the WGSA. For many phages, there are regions of the genome likely to be under-represented in the shotgun library, which results in more gaps in the shotgun assembly than predicted by the Poisson distribution. However, as phage genomes are relatively small, this increased number of gaps does not present an insurmountable impediment to using the WGSA. This chapter will focus on construction of a high-quality random library and sequence analysis of this library in a 96-well format. Techniques are described for the mechanical fragmentation of genomic DNA into 2 kb average size fragments, preparation of the fragmented DNA for shotgun cloning, and advice on the choice of cloning vector for library preparation. Protocols for deepwell block culture, plasmid isolation, and sequencing in 96-well format are given. The rationale for determining the total number of random clones from a library to sequence for a 50 and 150 kb genome is explained. The steps involved in going from hundreds of shotgun sequencing traces to generating contigs will be outlined as well as how to close gaps in the sequence by primer walking on phage DNA and PCR-generated templates. Finally, examples will be given of how biological information about the phage genomic termini can be derived by analysis of the *organization* of individual clones in the shotgun sequence assembly. Specific examples are given for the circularly permuted termini of *pac* type phages, the direct terminal repeats found in most T7-like phages, variable host DNA at either end as in the Mu-like phages, and the 5′ and 3′ overhanging ends of *cos* type phages. The end result of these steps is the entire DNA sequence of a novel phage, ready for gene prediction.

Key words: Shotgun sequencing, Poisson, phage genomics, *cos*, *pac*, genomic termini, undergraduate education.

Martha R. J. Clokie, Andrew M. Kropinski (eds.), *Bacteriophages: Methods and Protocols, Volume 2: Molecular and Applied Aspects, vol. 502*, © 2009 Humana Press, a part of Springer Science+Business Media
DOI 10.1007/978-1-60327-565-1_4 Springerprotocols.com

1 Introduction

On the surface, the sequencing of a phage genome appears to be a trivial endeavor—sort of an early morning project for any major sequencing facility. After all, most dsDNA phages are small, typically between 20 and 200 kb with only a few outliers with huge genomes *(1)*. In practice, though, the small genome sizes mean that the majority of time spent sequencing a phage genome is during the labor intensive initial library preparation steps and assembly finishing stages. Furthermore, phage genomes are "gappy," i.e., assemblages result in gaps at a higher frequency then predicted by Poisson calculation. There are many reasons for this, including toxic DNA sequences, non-random fragmentation, modified DNA, or toxicity of the encoded protein, the last of which is strongly affected by the choice of cloning vector. Additionally, phage genomic diversity, including genome organization is extremely high *(2)*. The increasing number of completely sequenced phage genomes has revealed an extraordinary diversity and complexity of the phage population and indicates that we do not know the most common phages for even the well-studied bacterial species *(2, 3)*. In contrast to viruses of animals and plants, when a new phage is isolated it is still reasonable to expect that the phage could exhibit an almost entirely novel arrangement of genes and there will almost certainly be large sections encoding proteins with limited or no recognizable homologs in the database. The purpose of this chapter is to provide the outline of steps and a standard set of protocols to generate a random, small insert library from phage genomic DNA with the goal of obtaining the complete sequences of a phage genome by a shotgun approach.

It is possible to perform most of the steps of library preparation and utilizing commercial kits. As well as usually saving time, commercial kits provide standardized reagents. It is also possible to perform all of the steps described here without using commercial kits. The techniques in and of themselves, however, are quite basic and generic molecular biology protocols will suffice *(4)*. Because most of the individual techniques are fundamental molecular biology protocols, the emphasis in this chapter is to describe the correct progression of steps required to obtain the genomic sequence. Most of the recommended protocols are based on commercial kits with only slight modification from the manufacture's protocol. The progression of steps and choice of recommended protocols described here were developed during the sequencing of 16 novel phages, primarily with hosts from members of the *Burkholderia cepacia*

complex (Bcep phages). These phages include virulent phages isolated from soil samples and several induced from lysogens *(5, 6)*. One rationale for sequencing these phages was purely didactic, that is they were sequenced by undergraduate students in order to teach WGSA theory and methods. Therefore, the recommended protocols are simple and robust enough for inexperienced users.

1.1 Outline of Steps in Phage Shotgun Genome Assembly

I. Shotgun library preparation:
   1. Isolate phage genomic DNA.
   2. Determine the size and purity of the phage DNA.
   3. Random fragmentation of the genomic DNA into 2 Kbp average size.
   4. Generate blunt ends on the fragmented DNA.
   5. Gel purify 2,000 bp average size fragments.
   6. Ligate DNA into blunt end, dephosphorylated vector.
   7. Transform ligation reaction into *E. coli* cells.

II. 96-Well format bacterial culture and plasmid purification:
   1. Inoculate 96 place deepwell blocks from the library sufficient to obtain eightfold random sequence coverage.
   2. Isolate plasmids in a 96-well format.

III. 96-Well format sequencing and reaction clean up:
   1. Sequence plasmids with left and right vector primers.
   2. Purify sequencing products, resolve on capillary sequencer.

IV. Shotgun sequence assembly, primer walking gap closure:
   1. End and vector trim sequences; assemble into contigs using sequence assembly software.
   2. Close gaps and resolve ambiguities by PCR and primer walking.

VI. Completing the genomic sequence:
   1. Determine the structure and sequence of the genomic termini.

2 Materials

2.1 Shotgun Library Preparation

Ten milliliters of a filtered 10^9–10^{10} pfu/ml lysate.

Nuclease Mix: 0.25 mg/ml RNAse A, 0.25 mg/ml DNAse I, 150 mM NaCl, 50% glycerol, store at −20° C

Phage Precipitant (20% w/v PEG 8,000, 1.76 M NaCl, store at room temparature)

SM buffer (100 mM NaCl, 8 mM MgSO4•7H_2O, 50 mM Tris, pH 7.5) Wizard DNA Clean Up kit (Promega)

5.4M guanidine thiocyanate

HydroShear device (GeneMachines, Genomic Solutions).

End repair enzymes (Lucigen).

Reagents and equipment for standard and pulse field gel electrophoresis (BioRad).
QIAquick Gel Extraction Kit (Qiagen).
CloneSmart HCKan Kit (provided with prepared pSMART HCKan vector, T4 DNA ligase, and 10G ELITE Electrocompetent Cells) (Lucigen).
Electroporator and 0.1 cm gap size electroporation cuvettes (Bio-Rad MicroPulser and cuvettes).
LB agar plates with 30 mg/L Kanamycin.
DMSO (>99.9% purity, Sigma).

2.2 96-Well Bacterial Growth and Plasmid Preparation

A 130 ml LB-Kan (30 mg/L) per 96 square-well culture block.
Sterile tooth picks, at least 100 per square-well culture block.
Shaking incubator with square-well culture block holders.
Centrifuge with swing out rotor with 96-well culture block carriers (IEC Centra).
96 place plasmid preparation kit (Qiagen R.E.A.L Prep 96).
96-well flat bottom plate with lid (Falcon).
Liquid handling robot (Qiagen BioRobot 3000).
100% Isopropanol and 70% ethanol.
Plastic sealing tape (Qiagen).

2.3 96-Well Format Sequencing and Reaction Cleanup

BigDye® Terminator v3.1 Cycle Sequencing Kit (100 μl/96-well plate) (ABI).
MicroAmp® Optical 96-well reaction plate (ABI).
MicroAmp® 96-well full plate covers (ABI).
BigDye® Terminator v1.1/v3.1 sequencing buffer (5X).
Eight place strip tubes.
Thermal cycler with 96-well plate capacity (ABI 2720).
75% Isopropanol and 70% Ethanol, made fresh.
PCR plate holders for centrifugation (USA Scientific).
Forward and reverse sequencing primers, 100 μl of 10 μm per 96-well plate.
Multi-channel pipettors, 10, 50, and 300 μl capacity.
Plastic sealing tape for 96-well plates (Qiagen).
Sequencing facility.

2.4 Shotgun Sequence Assembly, Gap Closure, and Final Sequence Assembly

Computer with sequence assembly software.
Primer design software (Primer3).
Sequencing reagents for single or strip tubes (ABI).
PCR reagents (ABI).
T4 DNA ligase (Promega).

3 Methods for the Shotgun Assembly of a Phage Genome

A schematic overview of WGSA is shown in **Fig. 4.1**.

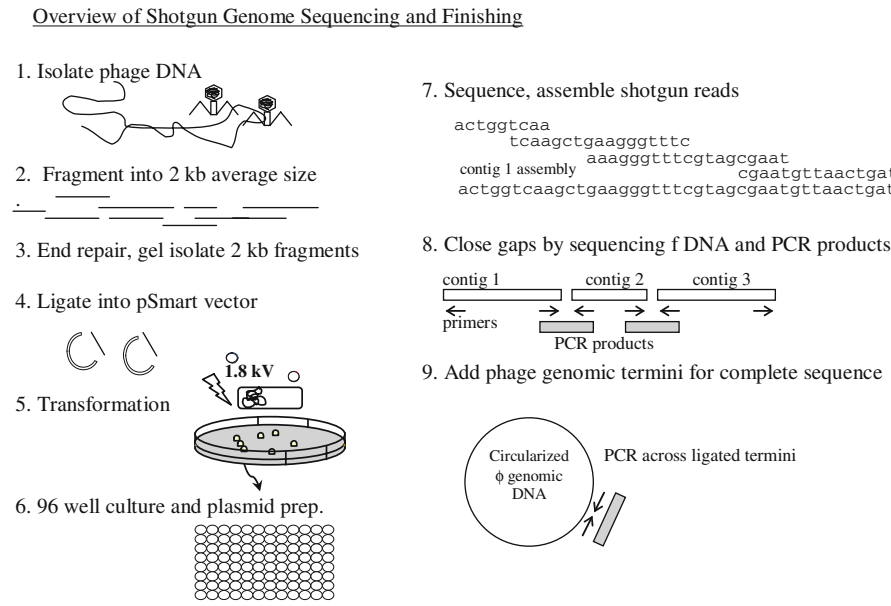


Fig. 4.1. Overview of shotgun sequencing and finishing 1. Purify phage genomic DNA free of bacterial host DNA. 2. Fragment randomly into 2 kb average length pieces. 3. End repair, size select. 4. Ligate into selectable vector, transform. 5. Pick clones at random and isolate plasmids from 96-well cultures. 6. Sequence inserts with right and left vector primers. 7. Match overlaps in fragments, generate consensus sequence. 8. Fill gaps and low quality sequence by sequencing phage DNA and PCR products. 9. Determine the genomic terminal structure and sequence.

3.1 Shotgun Library Preparation

3.1.1 Purify Phage Genomic DNA from 10 ml of 10^9–10^{10} pfu/ml Lysate (see Note 1)*

For many phages, it is possible to isolate DNA of sufficient purity directly from a filtered, high titer lysate by first nuclease treating the lysate to remove host nucleic acids, then precipitating phage particles by polyethylene glycol (PEG) precipitation and finally using a commercial kit designed for DNA purification and concentration. A kit that has given consistent results with *Burkholderia* phage is the Wizard DNA Clean Up kit (Promega).

1. To degrade bacterial host nucleic acids, add 40 μl Nuclease Mix to 10 ml of a 10^9–10^{10} pfu/ml filter sterilized (important) lysate in an Oakridge tube. Mix by inversion, incubate at 37°C for 30 min, followed by room temperature for 1 h.
2. To precipitate phage particles, add 7.5 ml Phage Precipitant, mix well by inversion. Incubate 30 min on ice or overnight at 4°C, mixing occasionally by inversion. Pellet phage particles by centrifugation at 12 K g for 20 min. Pour off supernatant and drain excess liquid off of the pellet by briefly inverting on paper towel.
3. To remove insoluble materials, resuspend the PEG/phage pellet in 0.5 ml SM buffer (100 mM NaCl, 8 mM $MgSO_4 \bullet 7H_2O$, 50 mM Tris, pH 7.5), transfer to 1.5 ml microcentrifuge tube. Pulse to maximum speed in microcentrifuge, transfer supernatant to a clean 1.5 ml tube. At this point, switch to using the Wizard DNA Clean Up kit

(Promega cat. # A7280) supplies and reagents, with the following modifications.

1. Uncoat phage DNA: Resuspend pellet by pipetting in 1 ml DNA Clean Up Resin from the kit (that has been warmed thoroughly to 37°C and has no guanidinium thiocyanate crystals). Pipet up and down gently to resuspend the pellet, then mix thoroughly by gentle swirling. The resin is provided in a solution of guanidinium thiocyanate which is the chemical that actually denatures phage proteins, releasing the phage DNA. The free phage DNA then binds to the resin.
2. Remove exopolysaccharides from sample (*this step greatly reduces clogging of the column*): The phage DNA is bound to the resin beads at this point. Transfer liquid and resin to a clean microfuge tube. Spin 5 min in microfuge (max speed). Pipette off ~2/3 of the liquid – leave behind the whole resin pellet with some liquid (the pellet is not very strong). Resuspend the pellet in fresh 1 ml 5.4 M guanidine thiocyanate.
3. Attach the column supplied with kit onto a 3 or 5 ml syringe.
4. Apply resin/phage DNA solution to column using a pipet. Use the syringe plunger to push this solution through the column.
5. Wash salts and proteins off of DNA: Remove the column from the syringe. Then remove plunger from syringe, then reattach the column. Add 2 ml 80% Isopropanol and push through column with plunger.
6. Dry column: This step removes the isopropanol. Remove column to a clean 1.5 ml centrifuge tube. Spin dry the column 5 minutes max speed in microcentrifuge.
7. Elute phage DNA off column: Transfer column to a clean microcentrifuge tube. Apply 80°C elution buffer onto the resin in the column, spin 1' to elute DNA. Repeat elution with a second aliquot of 80°C elution buffer.
8. Final DNA sample: Combine elutions, heat to 70°C for 10 minutes to inactivate any contaminating nucleases. This is the purified phage DNA. Useful yields range from 10 to 100 ng DNA/μl. Run a gel with 10 μl of the DNA. Store DNA at −20°C for long-term storage, in the refrigerator during periods of heavy use to avoid multiple freeze-thaw cycles.

3.1.2 Determine the Yield, Purity, and Size of the Phage Genomic DNA

Quantify the phage genomic DNA spectrophotometrically. The yield and quality should also be assessed by standard agarose gel electrophoresis of 100–250 ng of the DNA sample. The dominant phage DNA band typically migrates slightly above the 23 kb λ HindIII size standard. It is critical that the phage genomic DNA used for shotgun library preparation is predominantly free of

contaminating host DNA. Host DNA appears as a band that migrates very slowly, well above the phage DNA. There should also not be observable levels of RNA or degraded DNA that migrates below the 2 kb band of the λ HindIII digest. The length of the phage genomic DNA should be estimated by pulse field gel electrophoresis. Accurate sizing of the phage genome is required in order to determine how many plasmids should be isolated from the shotgun library and sequenced (this is explained in **Section 3.2**).

3.1.3 Fragmentation of the Phage Genomic DNA

Hydrodynamic shearing is the best method for producing randomly fragmented DNA (*see* **Note 3**). Bring 1–5 μg phage genomic DNA to 100 μL with sterile, filtered ultrapure H_2O. Spin sample in microfuge 10 k*g*, 1 min, transfer the sample to a new 1.5 ml microfuge tube. Use the GeneMachines HydroShear device (Genomic Solutions) to fragment this sample to 2,000 bp average size (20 cycles at a speed code of 7). Clean and use the HydroShear device according to the manufacturer's guidelines. Filter the wash solutions through a 0.45 μm filter to minimize clogging. A volume loss of 10% is typical; much more suggest a problem with the instrument.

3.1.4 End Repair the Sheared DNA

Hydro-sheared DNA contains a mixture of ends including 5′ and 3′ single-stranded overhangs. The ends must be converted into blunt ends prior to ligation into the vector. This process is variously termed polishing, blunt ending, or end repairing. An efficient commercially available end repair kit specifically optimized for this procedure is the DNATerminator End Repair Kit (Lucigen). This has the distinct advantage of not requiring optimization. Follow the manufacture's protocol.

Alternatively, a combination of T4 DNA polymerase and Klenow fragment can be used to generate blunt ends:

80 μL hydro-sheared DNA (at 5 to 50 ng/μl) sample
10 μL 10x Klenow Buffer
10 μL 2.5 mM dNTPs
1.5 μL T4 DNA Polymerase (NEB, 3 U/μL)
4.2 μL Klenow fragment (5 U/μL)
105.7 μl final volume

Mix gently but thoroughly by pipeting.
Incubate at room temperature (22 °C) for 40 min.
Heat inactivation of the end repair reaction: add EDTA to 10 mM; incubate at 75 °C for 20 min.

3.1.5 Gel Purification of the Hydro-sheared, End-Repaired Phage Genomic DNA

It is necessary to gel purify the 2 kb average size DNA fragments prior to ligation. Even if it is not visible on a gel, there are many small fragments (in the order of less than 100 bp) in the hydro-sheared mix. These small fragments will clone at a very high efficiency as compared to the larger DNA fragments and will dramatically reduce the average insert size of the library. A standard TBE agarose gel is sufficient for this step. All solutions and

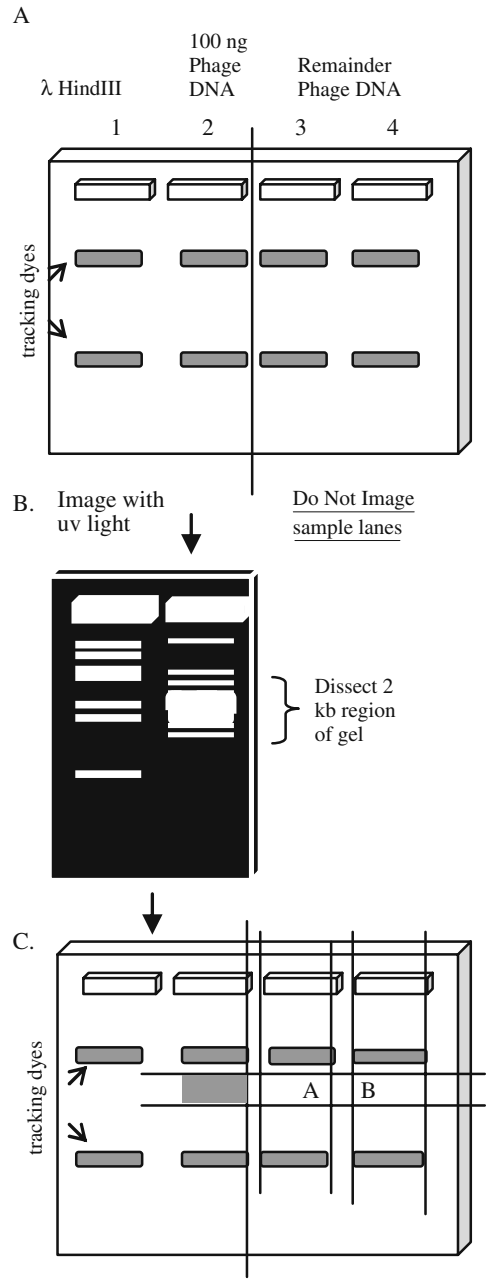


Fig. 4.2. Gel extraction the DNA to be cloned without UV irradiating **A**. Agarose gel loading order. Lane 1. DNA size standard (λ. HindIII). Lane 2. Aliquot of hydro-sheared, end-repaired phage genomic DNA (100 ng). Lane 3. Half of the remainder of hydro-sheared, end-repaired phage genomic DNA. Lane 4. Half of the remainder of hydro-sheared, end-repaired phage genomic DNA. Electrophorese the samples at 100 V for a short time (30 min to 1 hr). Stain the gel with ethidium bromide. Bisect the gel between Lanes 2 and 4 (in the *blank lane). **B**. Image the half of the gel with Lanes 1 and 2, excise the portion of the gel containing the desired fragment size. **C**. Re-align the bisected gel. Using the excised region in the first half of the gel as a template, excise the region of Lanes 3 and 4 that contain the desired fragment sizes. Transfer agarose plugs A and B to a 1.5 ml microfuge tube for DNA extraction.*

the gel casting apparatus need to be thoroughly rinsed and all gel solutions should be fresh in order to minimize the possibility of inadvertently cloning contaminant DNA. Do not expose the DNA to be cloned to short wave ultraviolet (UV) light. If possible, use long wave UV light to image the sample. If this is not possible, load, image, and dissect the gel as described in **Fig. 4.2** in order to perform size selection of the DNA to be cloned without exposure to short wave UV light.

3.1.6 Extracting the Hydro-sheared, End-Repaired DNA from the Agarose Gel Slice

Several companies manufacture kits designed for extracting DNA from agarose gels such as the QIAquick Gel Extraction Kit (Qiagen). Elute the DNA in 30 μl elution buffer. At this point, a diagnostic gel should be run with 5–15 μl of the purified DNA to confirm the success of the gel isolation procedure and to estimate the concentration of DNA. For alternatives to kits for gel isolation, *see* **Note 4**.

3.1.7 Ligation of End-Repaired DNA into pSMART-HCKan

A vector particularly well suitable for generating phage genomic libraries is pSMART-HCKan (Lucigen). In addition to lack specific promoters positioned to transcribe the insert, the pSMART vectors have transcription termination signals flanking the insert site to reduce transcription from cryptic promoters in the vector. These features allow for the stable cloning of even such extremely toxic phage proteins as holins. These vectors are purchased predigested and dephosphorylated and prepared for ligation. *See* **Note 5** for issues associated with choosing an alternative vector for phage genomic library construction.

Ligation of prepared phage insert DNA into pSmart-HC Kan vector (as per manufacture's protocol):

(X) μl hydro-sheared, end-repaired, size-selected phage DNA (100–500 ng).
(Y) μl H_2O.
2.5 μl 4× pSmart-HC Kan vector premix (contains buffer, ATP, and the prepared vector).
1 μl T4 DNA Ligase.
10 μl final volume.

Mix by gentle pipeting. Incubate at room temperature for 30 min to 2 h. Heat inactivate the reaction by incubation for 20 min at 70 °C (heat denaturing of ligation reactions is critical for transformation efficiency as it removes the enzyme from the DNA). The sample is ready for transformation.

3.1.8 Transformation of the Ligation Reaction

Electroporation or heat shock competent cells can be purchased or made (*see* **Note 6**). Electroporation competent *E. coli* 10G cells (*E. cloni* 10G genotype: F-*mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) ϕ80*dlac*ZΔM15 Δ*lac* × 74 *end*A1 *rec*A1*ara*D139

Δ(*ara, leu*)7697 *gal*U *gal*K *rps*L *nup*G λ-*tonA*) are provided with the CloneSmart HCKan Kit (Lucigen). Dilute the heat-denatured ligation mixture 1:1 with H_2O in order to reduce arching. Chill on ice before transferring 1 μl reaction to the aliquoted, chilled electrocompetent cells. Follow the guidelines for electroporation depending on the cells used. After the 1 h recovery, plate 50 μl of each transformation reaction onto the LB-Kan plates, incubate overnight at 37 °C. Store the rest of the transformation reaction at 4 °C. It is critical to perform the necessary controls (*see* **Note 7**).

The next day estimate the number of colonies on the plates. Plate out a volume of the remaining transformation reaction to obtain 100–200 well-separated colonies on five to eight more LB-Kan plates. The remainder of the transformation reaction can be stored at −80 °C following the addition of 9 μl DMSO (>99.9% purity, Sigma) per 100 μl cells.

3.2 96-Well Bacterial Growth and Plasmid Preparation

These procedures assume the use of the R.E.A.L. Prep 96 Plasmid Kit (Qiagen) on a BioRobot 3000 (Qiagen) liquid pipeting robot. Other kits and instruments, such as the Wizard MagneSil Plasmid Purification System (Promega) or the Biomek 2000 (Beckman Coulter), can be substituted. The R.E.A.L. Prep 96 Plasmid Kit can also be used without the BioRobot 3000 if a 1 ml volume, eight-channel pipettor, and a QIAvac 96 Vacuum Manifold are available.

Table 4.1
Relationship between fold coverage, number of clones sequenced, percentage of the genome obtained, and number of gaps in the final assembly for a 50,000 and 150,000 bp genome. Based on tables found at http://www.genome.ou.edu/poisson_calc.html

Genome size	Coverage	Number of sequencing reactions[a]	Number of plasmids isolated[b]	Total shotgun sequence (bp)	Percentage of Genome sequenced	Number contigs
50,000	1×	100	50	50,000	63	38
50,000	5×	500	250	250,000	99.4	4
50,000	8×	800	400	400,000	99.97	2
50,000	10×	1,000	500	500,000	99.995	2
150,000	1×	300	150	150,000	63	112
150,000	5×	1,500	750	750,000	99.4	11
150,000	8×	2,400	1,200	1,200,000	99.97	2
150,000	10×	3,000	1,500	1,500,000	99.995	2

[a] This assumes a 500 bp usable sequence read from each reaction.
[b] This assumes that all clones have a >1,000 bp phage DNA insert.

The number of 96-well plates of plasmids to isolate depends on the size of the phage genome, the coverage goal, the percentage of clones in the library with useable inserts, and the average sequence read length returned. For example, to obtain eightfold coverage of a 50 kb genome with 500 bp average usable sequence read lengths, ~400 clones (4× 96-well plates) from the shotgun library would need to be isolated and sequenced with left and right primers. *See* **Table 4.1** and **Note 2** for an explanation of the calculations used for shotgun genome sequencing.

3.2.1 96-Well Culture

Fill square-well culture blocks (provided with R.E.A.L. Kit) with 1.3 ml LB-Kan/well. Inoculate each well with a single, well-isolated colony using a toothpick, leaving the toothpicks in the wells to mark which has been inoculated. When all 96 wells contain a toothpick, remove, and discard toothpicks. Cover inoculated square-well blocks with the loose fitting lids provided or AirPore tape sheets (Qiagen) to allow for gas exchange during incubation. Incubate overnight at 37 °C shaking at least 270 rpm. The square-well culture blocks should be fitted either into holders designed for this purpose or wedged securely into wire baskets attached firmly to the rotating platform.

Next morning:

1. Pellet the bacteria at $1,500 \times g$ for 5 min on a swing out rotor designed to hold deepwell plates.
2. Drain the supernatant by pouring off the liquid and blotting dry on paper towels. Proceed to plasmid preparation immediately or cover with plastic sealing tape and store at −20 °C (good for at least 3 months).

3.2.2 Extract Plasmid According to the Kit/Robotic Pipeting Tool Combination Available

Component	per reaction	MM × 100
ddH_2O	3.5 µl	350 µl
5 × RB	1 µl	100 µl
10 mM primer	1 µl	100 µl
BigDye	1 µl	100 µl
DNA	1 µl	*
Total	*7.5 µl*	*650 µl*

The final volume of plasmid DNA from a 1.3 ml overnight culture should be 100 µl for optimum sequencing. Add additional H_2O to the eluted DNA as needed (*see* **Note 8**). Transfer the plasmid to a 96-well flat bottom plate with lid. For storage, seal the plate well with plastic sealing film, replace the 96-well flat bottom plate

lid, and store at −20 °C. After thawing the plate, pulse liquid to the bottom of the wells by centrifugation to $1,000 \times g$ and back.

3.3 96-Well Format Sequencing

3.3.1 96-Well Sequencing Reaction Mix Preparation

For each 96-well plasmid prep plate, make the following master mix that contains all components except the template (*see* **Note 9**). Sequence all plasmids with both left and right primers so there will be 192 sequencing reactions for each 96-well plate of plasmid isolated.

Make up the master mix (MM) in 1.5 ml microfuge tube on ice. Add the components in the order listed (i.e., add the BigDye to the MM last) pipet up and down gently to mix. Transfer 81 µl of MM into each well of the eight place strip tube placed in a cold block. Dispense 6.5 µl MM into the very bottom of each well of MicroAmp Optical 96-Well Reaction Plate (ABI) (placed in a cold block), using a multi-channel pipette. Visually confirm that MM has been added to each well. Add 1 µl of template DNA from the plasmid plate directly into the MM at the bottom of each well, pipet up and down four times to mix. It is important to confirm that the orientation of the plasmid plate is the same as the sequencing reaction plate (i.e., plasmid A1 goes into reaction well A1) and to change tips between each sample. Cover the reaction plate with a MicroAmp 96-Well Full Plate Covers (ABI). Place in 96 position thermal cycler and start the run:

Sequencing reaction conditions with pSmart-HCKan left and right sequencing primers (*see* **Note 9**):
96 °C 1′
96 °C − 50 °C − 60 °C, 10″ − 5″ − 4′ for 99 cycles.
4 °C soak.

3.3.2 Purification of Sequencing Reaction Products

During removal of unincorporated nucleotides, it is critical that all reagents and centrifugations during sequencing reaction purification are performed at room temperature.

1. Add 30 µl (= 4× reaction volume) of 75% isopropanol to each well, mix by pipeting four times.
2. Cover plate with plastic sealing tape, place in plate centrifuge adaptor.
3. Centrifuge $2,500 \times g$ at least 2 hr (longer is fine).
4. Decant isopropanol by removing tape and inverting plate onto paper towels, blot off excess liquid. Do not right plate at this point.
5. To drain excess liquid, place inverted plate onto pad of paper towels in the plate carrier; pulse to 1,000 rpm and down.
6. Right plate, return to plate adaptor.
7. To each well, add 100 µl of 70% ethanol, do not mix or pipet, cover with plastic sealing film.
8. Centrifuge $2,500 \times g$ 30 min.

9. Decant ethanol by removing tape and inverting plate onto paper towels, blot off excess liquid. Do not right plate at this point.
10. Place inverted plate onto paper towels in the plate carrier; pulse to 1,000 rpm and down (to dry pellet).
11. Sequencing reaction products are now purified away from unincorporated BigDye and are in a pellet in the bottom of each well. At this point, either cover the plate with sealing tape or resuspend pellet in solution as dictated by the sequencing center (sequencing facilities vary in requirement at this step).
12. Analyze sequencing products on capillary sequencer.

3.4 Shotgun Sequence Assembly, Primer Walking Gap Closure, and Final Sequence Assembly

This requires the use of sequence trimming and assembly programs such as Phred/Phrap/Consed *(7)* although even simple programs such as Sequencher (GeneCodes) can be used. Phred/Phrap/Consed is a powerful program suite that automates many of the processes. Sequencer, however, is intuitively easy for students to use and manual editing provides students with the opportunity to learn how sequences are assembled. The steps involved in trimming the raw sequences and assembling into contigs include trimming the 5′ and 3′ ambiguities and vector sequences from each trace. All low quality or vector only sequences should be removed from the project. Contigs are then assembled from the shotgun sequences. The resulting contigs are edited to resolve ambiguities and to mark areas of low quality or low coverage sequences that require re-sequencing. Orientation of the end reads in the project provides support for the assembly, that is the left–right traces from the same plasmid should face each other and be no more than 3 kb apart. It is helpful to remember that the random shotgun clones are not directional, i.e., the left–right designation reflects the orientation in the vector not the phage assembly. This generates the random shotgun assembly of the phage genome (*see* **Note 11**).

3.4.1 Generating Shotgun Assembly 1

3.4.2 Primer Walking for Gap Closure and Low Quality Sequence Validation

Even at greater than 8× coverage, many phage genome assemblages will have more gaps than predicted. For example, when the projects were analyzed for six *Burkholderia* phage shotgun assemblages (sequenced to 7× coverage or greater), the number of contigs ranged from 1 to 15 with four having four or less contigs. Additionally, while the average coverage will be high, there will be individual regions that are not unambiguously sequenced on both strands (the minimum requirement for quality sequence). The initial strategy to close gaps would be to identify and re-sequence individual clones from the shotgun library predicted to span the gaps. Often, though, the ends of contigs are closed, that is there are no clones than

span gaps. When all potentially useful clones have been re-sequenced, then primers should be designed to sequence the phage DNA directly or to sequence PCR products. Primers should be positioned no closer than 75 or further than 400 nucleotides from the end of a contig. An excellent web site for primer design is Primer3 (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). In order to give the program the opportunity to select the optimal sequencing primer, paste in sequences located between 400 and 75 bases away from the region to be re-sequenced. A similar approach is used to design primers to re-sequence ambiguities with individual library clones being the initial choice followed by sequencing phage DNA directly or PCR products. If phage DNA is to be sequenced directly, use the same reaction described for sequencing library clones except omit the H_2O and add 4.5 μl of phage DNA to the 7.5 μl sequencing reaction. Phage lysate can usually be used directly in a PCR, but the DNA needs to be purified for direct sequencing. After assembling sequences derived in these manners, the project should be in a single contig. To finish the sequence, however, the phage genomic termini need to be elucidated.

3.5 Finishing the Genomic Sequence

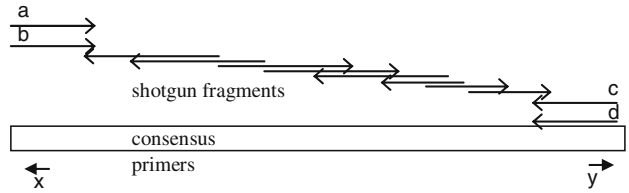
The genomic sequence is not complete until the terminal sequences are added. The terminal sequences cannot be known unless the structure of the genomic termini is determined. It is outside the scope of this chapter to provide a comprehensive description of phage genomic termini organizations. However, it is often possible to establish the phage genomic terminal structure and sequence by a combination of analysis of the organization shotgun library clones, sequencing phage DNA directly, and sequencing PCR products of ligated phage genomic DNA. At this point, analysis of the encoded proteins should have been initiated and may provide insight as to the predicted terminal structures. If the gene for the terminase large subunit has been identified, it is also useful to compare its amino acid sequence with that of known terminase large subunit sequences *(8)*. While not comprehensive, specific examples are given for five categories of phage genome terminal organizations: circular permutation of a *pac* type or random circular permutation type phage, 5′ overhang cohesive (*cos*) ends, 3′ overhang cos ends, direct terminal repeats, and variable host sequences.

3.5.1 If the Phage Shotgun Assembly Results in a Circular Map, the Phage Is Likely to Have *pac* Ends

Restriction digests of phage genomic DNA or restriction digests in combination with Southern analysis can be used to map the *pac* site which is detected as a submolar fragment that frequently maps near the terminase coding region *(8)*. Some phages, however, have highly circularly permuted genomes with no obvious

3A. Differentiating 5'and 3'overhanging *cos* ends

a. Schematic of *cos* phage shotgun assembly



b. Steps in obtaining 5'and 3'overhangs

1. Ligate phage genomic DNA
2. PCR ligated genomic DNA with end primers x and y
3. Sequence the PCR product
4. Sequence the phage genomic DNA directly with the same primers
5. Make a new contig consisting end clones a,b,c,d from the shotgun assembly and the sequence of the PCR product and the phage DNA.

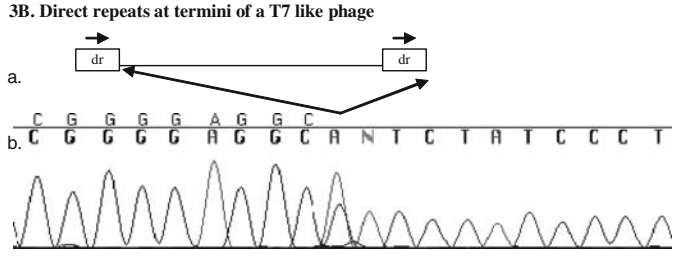
c. Assembly of sequences from 5'overhang *cos* ends

```
1. CGAAAAAACGGGAAATCGGGGCTCTGCGCCCT   shotgun clone c
2. CGAAAAAACGGGAAATCGGGGCTCTGCGCCCT   shotgun clone d
3. CGAAAAAACGGGAAATCGGGGCTCTGCGCCCT     ϕ DNA/primer y
4. CGAAAAAACGGGAAATCGGGGCTCTGCGCCCTCCCGCCATTCGCGCCTCG
5.    ϕ DNA/primer x   GCTCTGCGCCCTCCCGCCATTCGCGCCTCG
6.  shotgun clone a   GCTCTGCGCCCTCCCGCCATTCGCGCCTCG
7.  shotgun clone b   GCTCTGCGCCCTCCCGCCATTCGCGCCTCG
```

d. Assembly of sequences from 3'overhang *cos* ends

```
1. TGATGGCCGCCTCGGCCTAGAC   shotgun clone a
2. TGATGGCCGCCTCGGCCTAGAC   shotgun clone b
3. TGATGGCCGCCTCGGCCTAGAC     ϕ DNA/primer  x
4. TGATGGCCGCCTCGGCCTAGACCGCACGTTCCCCTCACGCGCAGAAAATTTT
5.      shotgun clone c             CTCACGCGCAGAAAATTTT
6.      shotgun clone d             CTCACGCGCAGAAAATTTT
7.      ϕ DNA/primer y               CTCACGCGCAGAAAATTTT
```

3B. Direct repeats at termini of a T7 like phage



3C. Variable host DNA flanking the termini of a Mu-like phage

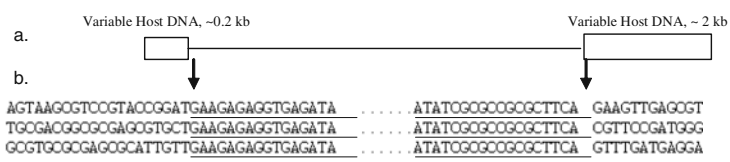


Fig. 4.3. (continued)

pac site *(9)*. In this case, the genomic sequence is generally linearized for annotation purposes.

3.5.2 If the Shotgun Assembly Produces a Linear Map with Defined Ends, the Phage Probably Has *cos* Ends

If the phage has cohesive (*cos*) ends, the left and right ends of the assembly will be comprised of numerous inward facing clones that, upon close inspection of the trimmed sequences, start at the same position. A combination of analysis of the individual clone, directly sequencing off the ends of the phage genome, and sequencing PCR products derived by amplification through the ligated junction of circularized genomic DNA is required to distinguish between 5′- and 3′- overhanging *cos* ends, as described in **Fig. 4.3A**.

3.5.3 If the Phage Is T7-Like, it will Probably Have Long Terminal Direct Repeats

Methods to define the boundaries of the several hundred base pair long repeats in T7-like phages can be particularly problematic, in part due to the lethality of the sequences at the left end

Fig. 4.3. (continued) Determining phage genomic termini. (A) Distinguishing 3′ and 5′ overhanging cos ends **a**. Schematic of a shotgun assembly of a *cos*-type phage genome. The middle bar represents the consensus line. Arrows indicate the position and sequence direction of individual clones in the assembly with a, b, c, and d representing shotgun library terminal clones. The location and orientation of primers *x* and *y* are indicated. Note that the ends of the assembly consist of several independent shotgun clones with an inward orientation. Manual editing of individual end clones should result in all of the sequences present in the shotgun clones. **b**. Outline of steps required to obtain the *cos* termini sequences from the assembly. **c**. Assembly of sequences from the genomic termini of a 5′ overhang *cos* phage. 5′ overhanging *cos* ends are filled in during the end repair reaction. The 5′ overhang sequence will be present in clones in the shotgun sequence (Lanes 1, 2, 6, and 7). As the 5′ overhangs are complementary, this will appear as short direct repeats at the ends of the consensus. This sequence is also present when phage DNA is directly sequenced with primers x and y (Lanes 3 and 5). The overlapping sequence (GCTCTGCGCCCT) is the short direct repeat at the genomic termini. No new sequence is obtained from sequencing the PCR product of ligated phage DNA (Lane 4). **d**. Assembly of sequences from a 3′ overhang *cos* phage. 3′ overhanging *cos* ends are removed during the end repair reaction. The 3′ overhang sequence will not be present in clones from the shotgun assembly (Lanes 1, 2, 6, and 7) nor will the sequence be present when phage DNA is directly sequenced with primers *x* or *y* (Lanes 3 and 7). The new sequence (CGCACGTTCCC) present in the PCR product of ligated phage DNA (Lane 4) is present at both genomic termini, forming short direct repeats. (**B**) Direct repeats at termini of a T7-like phage. **a**. Schematic of T7-like phage genomic DNA showing the direct repeats (dr) that flank the unique coding sequence. The arrows indicate primers that anneal in the direct repeats. **b**. Chromatogram resulting from sequencing T7-like phage genomic DNA with a primer located inside the direct terminal repeat. Notice at the point of the end of the direct repeat (the double peak at position 10), the peak intensity is reduced to ~50%. (**C**) Variable host DNA flanking the termini of a Mu-like phage. **a**. Schematic of Mu-like phage genomic DNA showing the variable host sequences that flank the unique coding sequence. **b**. Alignment of shotgun clones corresponding to the ends of the phage genomic DNA will have the consensus phage genome sequence and host sequences derived from regions throughout the host genome.

of the phage genome *(10)*. One method is to directly sequence the phage DNA with leftward and rightward primers located in the direct repeat (**Fig. 4.3B**). At the boundary of the direct repeat, the chromatogram should drop in intensity, which corresponds to a 50% decrease in template at the junction boundary, although this feature may be more pronounced in one direction than the other.

3.5.4 The Genomic End Clones from a Mu-Like Phage Shotgun Assembly Will Be Chimeric

In this case, the sequences at either end of the genomic contig will consist of sequences possessing the defined phage termini and a random fragment of host DNA that is different in every independent clone. This is particularly striking and unmistakable (**Fig. 4.3C**) *(6)*.

4 Notes

1. If lambda (48.5 Kbp) is used as an example, 2×10^{10} phage particles contains 1 μg DNA *(4)*. Therefore, titers of 10^9–10^{10} pfu/ml are required to obtain sufficient DNA for all of the steps including running diagnostic gels, sizing gels, and to account for losses at each purification step.
2. Deciding how many plasmids isolate and sequence: Poisson distribution. The decision of how many plasmids to isolate and sequence is based on Poisson distribution *(11,12)*. These values represent an ideal based on the random nature of library generation. As phages have a host of characteristics that preclude truly random library generation, these values are the ideal. In principle, there is information to be derived in understanding why a phage shotgun assembly does not match the random ideal. For example, it is possible that nonrandom breaks are due to specific single-strand nicks in the genomic DNA backbone as observed in phage T5 *(13)*. Also, large unclonable regions may exist due to the presence of toxic DNA sequences such as strong host promoters present at the left end of some T7-like phages *(10)*.

   For 50 and 150 kb genome sizes, it is reasonable to obtain eightfold coverage. Fold coverage is the total random phage genomic sequence obtained divided by the length of the target genome. For example, if 400 randomly chosen plasmids from the shotgun library were sequenced with two separate primers (= 800 sequencing reactions), and an average of 700 bp of usable sequence was obtained from each reaction, there would be 800 reactions $\times 700$ bp/reaction $= 560,000$ bp of random shotgun sequence. If the target phage genome was 50,000 bp, then

the coverage would be $560,000/50,000 = 11.2$-fold. If the target phage genome was 150,000 bp, then the coverage would be $560,000/150,000 = 3.7$-fold. As coverage increases, the number of gaps in the final assembly is predicted to decrease, thus reducing the number of labor intensive primer walking steps needed to finish the sequence. As sequence read length increases, the total number of plasmids that need to be isolated and the total number of sequencing reactions required to obtain a specific coverage decreases, which dramatically lowers the final cost of the project. A small average insert size or poor sequence quality will increase the number of clones that need to be isolated and sequenced, thus increase the final cost. **Table 4.1** describes the relationship between fold coverage, number of clones sequenced, percentage of the genome obtained, and predicted number of gaps in the final assembly for a 50,000 and 150,000 bp genome.

3. The method of choice for random DNA fragmentation is hydro-shearing as it provides consistent results for a large range of DNA lengths and volumes and requires no calibration *(14)*. However, it does require specialized equipment. Alternative methods include nebulization and sonication *(15)*. An enzymatic method is partial DNAse I treatment in the presence of Mn^{2+} *(16)*. These methods all require more starting genomic DNA as they need to be calibrated.
4. There are protocols available for those who do not want to use a gel isolation kit. Common methods include binding to glassmilk *(17)* or freeze-squeeze *(18)*. Introduction of ligation reaction inhibitors are a potential drawback to most methods. As a control that the method used for gel isolation does not introduce ligase inhibitors, these procedures should be tested by purification and ligation of HincII digested DNA.
5. The choice of cloning vector will significantly affect the level of random coverage obtained. This is because many phage-encoded proteins are toxic to the cell when expressed even at extremely low levels. Most vectors exhibit some level of constitutive expression of the insert whether from background expression levels inherent in most inducible promoters, as part of improperly terminated transcripts generated from expression of plasmid borne genes (such as the antibiotic resistance gene) or from cryptic promoters. Most common vectors such as pUC19, and pBlueScript are extremely poor choices for phage genomic library preparation, as is any expression vector or any vector with blue/white screening for inserts.

   The vector used for generating the library needs to be linearized with a blunt end restriction enzyme (i.e., HincII or

SmaI) followed by dephosphorylation with either calf intestine or shrimp alkaline phosphatase. It is then necessary to check the completeness of the digest and phosphatase reaction by setting up control ligation reactions with and without λ HincII insert as well as a no ligase control, and then performing a transformation identical to the one that will be used under the experimental conditions with the three reactions. There should be >250 colonies with the λ HincII insert, <25 colonies with the no insert ligation, and no colonies with linearized vector only.

6. *Choice of transformation method*: Heat shock and electroporation are the two most common transformation methods. Electroporation requires an electroporator but it provides the highest number of transformants, with efficiencies $>10^{10}$ (transformation efficiency numbers refer to the number of colony forming units per microgram of supercoiled control DNA). In contrast, heat shock requires no special equipment, however efficiencies of 10^7 are typical. If there is no electroporator available, then an easy and efficient protocol for generating heat shock competent cells is provided by Inoue *(19)*. The key to making high efficiency heat shock (or electroporation) competent cells is to grow the cells to early mid-log phase, chill the cells on ice before processing, and to keep **all** components (i.e., buffers, centrifuge tubes, centrifuges, and rotors) pre-chilled to 4 °C during all steps. If heat shock cells are used, do not dilute the ligation reaction.
7. Ligation and transformation efficiency controls should be performed. The positive control ligation is the ligation of λ HincII digested DNA into the pSmart vector. The negative control is the ligation reaction with no insert added. These should be heated, diluted, and electroporated into cells at the same time as the phage genomic DNA sample. Additional transformation controls include positive transformation control with 10 pg of supercoiled pUC19 DNA and a negative control of no DNA added to the transformation reaction. Ideal results after plating 50 μl of the transformation reaction would be: pUC19 transformation control, >500 colonies; λ HincII and experimental phage genomic DNA ligations, >250 colonies; no insert ligation, <25 colonies; no DNA transformation, no colonies.
8. At this point, it is useful to run a gel of 5 μl of each plasmid preps from at least one entire 96-well plate per library in order to determine the ratio of clones with insert to clones with small or no insert. Use gel boxes and combs designed to be loaded with the multi-channel pipette. As the pSmart vector is 1.9 kb and the average insert size should be 2 kb,

there is no need to digest the plasmid prior to analysis or to run the gel very long.

9. If different primers are used, the annealing temperature for sequencing should be 5 °C lower than the melting temperature of the primer.
10. At this point, calculate the actual coverage obtained. To do this, determine the number of sequences in the assembly and the average sequence read length (select 50 individual sequence reads from the assembly at random and determine their average size). The product of these will be the total number of base pairs of good quality phage shotgun DNA sequence. Divide this number by the phage genome size in base pairs to give the actual coverage. Compare the number of contigs at this point to the predicted number of contigs in **Table 4.1** to see if the phage assembly fits a Poisson distribution.

References

1. Brussow H, Hendrix RW. Phage genomics: small is beautiful. Cell 2002;108(1):13–6.
2. Hendrix RW. Bacteriophage genomics. Curr.Opin.Microbiol. 2003;6(5):506–11.
3. Ackermann HW. Bacteriophage observations and evolution. Res.Microbiol. 2003; 154(4):245–51.
4. Sambrook J, Fritsch EF, Maniatis T. Molecular Cloning: A Laboratory Manual. 2ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory; 1989.
5. Summer EJ, Gonzalez CF, Boomer M, Carlile T, Embry A, Kucherka AM et al. Divergence and mosaicism among virulent soil phages of the *Burkholderia cepacia* complex. J.Bacteriol. 2006;188(1):255–68.
6. Summer EJ, Gonzalez CF, Carlisle T, Mebane LM, Cass AM, Savva CG et al. Burkholderia cenocepacia phage BcepMu and a family of Mu-like phages encoding potential pathogenesis factors. J.Mol.Biol. 2004;340(1):49–65.
7. Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing. Genome Res. 1998;8(3):195–202.
8. Casjens SR, Gilcrease EB, Winn-Stapley DA, Schicklmaier P, Schmieger H, Pedulla ML et al. The generalized transducing Salmonella bacteriophage ES18: complete genome sequence and DNA packaging strategy. J. Bacteriol. 2005;187(3):1091–104.
9. Thomas CA, Jr., MacHattie LA. Circular T2 DNA molecules. Proc.Natl.Acad.Sci.U.S.A 1964;52:1297–301.
10. Scholl D, Kieleczawa J, Kemp P, Rush J, Richardson CC, Merril C et al. Genomic analysis of bacteriophages SP6 and K1-5, an estranged subgroup of the T7 supergroup. J.Mol.Biol. 2004;335(5):1151–71.
11. Port E, Sun F, Martin D, Waterman MS. Genomic mapping by end-characterized random clones: a mathematical analysis. Genomics 1995;26(1):84–100.
12. Lander ES, Waterman MS. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 1988;2(3):231–9.
13. Wang J, Jiang Y, Vincent M, Sun Y, Yu H, Wang J et al. Complete genome sequence of bacteriophage T5. Virology 2005;332(1): 45–65.
14. Oefner PJ, Hunicke-Smith SP, Chiang L, Dietrich F, Mulligan J, Davis RW. Efficient random subcloning of DNA sheared in a recirculating point-sink flow system. Nucleic Acids Res. 1996;24(20):3879–86.
15. Roe B.A., Crabtree JS, Khan AS. DNA Isolation and Sequencing. John Wiley & Sons; 1996.
16. Anderson S. Shotgun DNA sequencing using cloned DNase I-generated fragments. Nucleic Acids Res. 1981;9(13):3015–27.
17. Boyle JS, Lew AM. An inexpensive alternative to glassmilk for DNA purification. Trends Genet. 1995;11(1):8.
18. Tautz D, Renz M. An optimized freeze-squeeze method for the recovery of DNA fragments from agarose gels. Anal.Biochem. 1983;132(1):14–9.
19. Inoue H, Nojima H, Okayama H. High efficiency transformation of Escherichia coli with plasmids. Gene 1990;96(1):23–8.

Chapter 5

PCR and Partial Sequencing of Bacteriophage Genomes

Martha Clokie

Abstract

PCR is a quick and effective way of identifying the presence and 'affiliation' of bacteriophages, or phage-encoded genes from environmental samples, bacterial cells or purified viruses. The limitations are that you have to know what you are looking for in order to find it. Although the bacteriophage world does not have the advantage of a conserved gene, present in all members, there are many phage genes that do show nucleotide conservation even between phages which infect fairly divergent taxa. As more sequence data become available through both metagenomic approaches and the sequencing of complete bacteriophage genomes, PCR primers can be further refined and thus it should be an increasingly useful tool for bacteriophage biology.

Key words: PCR, amplification, template, dNA, sequencing, primer.

1 Introduction

This chapter is not designed to be a panacea to all PCR and sequencing-related problems in bacteriophage biology, however it does systematically cover the main phage-specific issues regarding the technique. For the PCR neophyte or even beginner, the following website has very useful information: http://www.horizonpress.com/pcr/.

To characterise a newly isolated bacterial strain, one of the first tools used is the polymorphism of a universal marker, typically the sequence of the 16S ribosomal DNA. This gives information of the phylogenetic position of the bacterial strain. All bacteria have ribosomes and thus ribosomal genes which fortunately make good markers due to having highly conserved regions, thus 'universal' primers can be used to amply them and highly variable regions which can be used to distinguish between bacterial strains.

Martha R. J. Clokie, Andrew M. Kropinski (eds.), *Bacteriophages: Methods and Protocols, Volume 2: Molecular and Applied Aspects, vol. 502*, © 2009 Humana Press, a part of Springer Science+Business Media
DOI 10.1007/978-1-60327-565-1_5 Springerprotocols.com

Unfortunately, bacteriophages have no such gene that is common to all representatives. Even when a gene may be common to all three phage families, e.g. the major structural genes, the sequences are so varied that an alignment and consequently a phylogenetic analysis is not possible (**Volume 2 Chapter 6**). One method to estimate viral diversity is to use metagenomic sequencing (**Chapter 23**) followed by custom designed arrays to probe known viral sequences in a range of environments. However, this scale of project is still expensive in terms of labour and cost. Often it is useful to just get an idea of what family a virus belongs to, or whether it possesses a particular gene of interest. Alternatively one may be investigating the presence of phages or phage-encoded genes as prophages or in bacterial communities. For example, phage-encoded toxins contributing to pathogenesis in cholera, diphtheria, enterohemorrhagic diarrhoea, and *Staphylococcus aureus* were successfully screened for using PCR-based approaches *(1)*. For many small-scale focused projects, PCR is still the preferred technique.

The first step in bacteriophage PCR is obtaining suitable high-quality template from which to amplify genes. Fortunately, in most cases, PCR requires very little DNA (~1–10 ng of phage DNA per reaction), however, there are cases where more is required (such as direct sequencing). DNA may be extracted from environmental samples (**Chapter 2**), phage lysate prepared from either a liquid culture or scraped plates (**Chapter 23**), infected cells or uninfected cells if prophages are being looked for. cDNA is also a good template for PCR, in which case good quality RNA has to be either extracted from infected cells or isolated from a bacterial community (**Volume 2 Chapter 9**). The template can also be diluted high titre phage lysate or DNA removed from community DNA extracted from agarose gels.

One method of increasing the amount template for downstream applications, such as sequencing is via the use of phage-derived enzymes, e.g. Genomiphi from GE Healthcare. This enzyme (the DNA polymerase from *Bacillus* phage φ29) is particularly useful for samples, where it is difficult to obtain large quantities of DNA, since it converts nanograms of DNA to micrograms of DNA overnight. The product can then be suitably used for direct sequencing, PCR or library preparation (**Volume 2 Chapter 1**). Direct sequencing is particularly useful, for example, where the presence of a gene in a novel genome has been detected using PCR, but it is not known what genes lie to either side. It is, therefore, possible to design primers to walk out from the known gene and to walk out using the total amplified phage genome as a template *(2)*.

The biggest challenge in bacteriophage PCR and sequencing is the choice of gene to be amplified and the consequent design of primers. The gene of choice is obviously dependent

on the question being addressed. For questions of diversity, the best understood bacteriophage 'genus' is the 'T4-like' viruses in the family *Myoviridae*, where primers have been designed for the major capsid proteins gp23 and gp20 and for the gene which encodes DNA polymerase (gp43). These have been predominantly used to study diversity in cyanophages *(3, 4, 5)*, but they have also been effective in oceanic viruses in general *(6)*. Caution must always be taken, however, in the interpretation of author's claims to virus primers being 'universal' as they may not be ubiquitously so. Even genes which are common to all 'T4-like' myoviruses are widely divergent *(7)*.

No such similar environmental screens have been carried out for oceanic *Siphoviridae* and *Podoviridae*, which likely reflects the paucity of knowledge of these phage sequences. As more sequences, data become available both through sequencing metagenomes and individual phage genomes, probing the 'phageome' with PCR-based approaches will become easier. Similarly viruses which infect archea have been less studied in many ways than phages of eubacteria and due to the high diversity of genes that have been identified, no such environmental screens of diversity based on PCR have been attempted. When sufficient study has been performed on a group of phages in a particular environment, sequence data can be generated which then lends itself to further study by PCR. An example of this is in the dairy industry where bacteriophages are a major economic problem. In *Lactococcus*, most phages belong to one of the three major groups of *Siphoviridae (8)*. This allows multiplex PCR primers to be designed for a relatively rapid screen of dairy plants. However, recent more detailed analysis has identified many novel groups even within this narrow environment that cannot be detected with PCR *(9)*.

When working with individual bacteriophages, electron microscopy can be used to establish the family to which they belong (**Chapter 12**). Then at least it is possible to establish, for example, whether the phage was a myovirus and whether to expect a gp20 product from a PCR. Although the morphology of viruses does not always reflect their genetics, if the 'genus' within the phage family can be established (**Chapter 12**), there is a higher chance of using this information to design appropriate primers.

If the aim of the project is to identify temperate bacteriophages, then primers specific to integrases may be a suitable target although it is worth remembering how varied such sequences can be *(10)*. The identity of the bacteriophage family is not so useful when screening for bacteriophage-encoded genes acquired from their bacterial hosts. For example, cyanophage acquired photosynthesis genes may be amplified equally successfully with the same primers from either podoviruses or myoviruses *(11)*.

If no sequence data is available for the bacteriophage/s being studied, then it may be possible to design primers from closely related bacteriophages. However as sequences can be so varied, even with significant degeneracy often this may be of little use. In an ideal world, one could completely sequence the entire genome of the bacteriophage (**Volume 2 Chapter 1**), however in reality, this is currently outwith the scope of most projects and budgets. The presence or absence of particular genes may be determined using Southern blotting. Cold and radiolabelled RFLPs (**Volume 2 Chapter 1**) may be done in parallel and the appropriate fragment of gene from the cold version can be cloned and sequenced. If more than one bacteriophage has been isolated for a new organism, it may be useful to determine the genes they have in common as if the phages are in the same family, they are likely to share structural genes. It may be the case that even hybridisation approaches fail and one solution is to take proteomic approach to identify the major structural proteins (**Chapters 18 and Volume 2 Chapter 14**). This will allow degenerate primers to be designed and sequence data generated. Although this approach may seem cumbersome, it has been successful when trying to obtain sequence data for poorly characterised bacteriophage taxa *(12)*.

2 Materials

All solutions, plastic and glass wear and equipment should be clean and sterile. PCR reagents and DNA should be kept at −20 °C. All other solutions should be kept at room temperature unless stated otherwise.

2.1 Target DNA (Template) for PCR

This is obtained as described below.

2.2 Target DNA (Template) for Sequencing

1. Nanogram quantities of DNA template.
2. Commercially available kit for amplifying circular DNA, e.g. Genomiphi (GE Healthcare; http://www.gehealthcare.com/formerly Amersham Bioscience).

2.3 Primers

These can be purchased from any oligo synthesising unit—e.g. Sigma Genosys (The Woodlands, Texas; http://www.sigmaaldrich.com/Brands/Sigma_Genosys.html), Invitrogen Corp. (Carlsbad, CA; http://www.invitrogen.com/) and many small companies.

2.4 General Reagents for PCR

1. Ultra-pure H_2O to 50 μl.
2. PCR polymerase enzyme with corresponding PCR buffer (commercially available).

3. 5 µl of 2 mM dNTP solution (commercially available from PCR enzyme producer).
4. Primers: 10 µM working solution in ultra-pure H_2O (*see* **Note 1**).
5. 1 ng – 1 µg/µL template DNA.

2.5 General Reagents for Gel Electrophoresis Visualisation

1. Agarose.
2. Ethidium bromide (50 µg/ml).
3. Loading buffer (6X concentration, to make 100 ml): add 93.6 ml of glycerol to 153.4 ml of water, 3 ml of 0.3 M EDTA, 0.3 g of bromophenol blue and 0.3 g of xylene cyanol.
4. TAE running buffer (50X stock solution): 24.2% (w/v) Tris–HCl (pH 7.5), 5.71% (w/v) acetic acid, 3.72% (w/v) EDTA. $2H_2O$ (adjust to pH 8 with HCl).
5. DNA size concentration ladder (e.g. New England Biolabs; Ipswich, MA; http://www.neb.com/or Fermentas Inc; Hanover, MD; http://www.fermentas.com/).

2.6 Equipment

1. Thermocycler or PCR machine. Any make will do. If using an established protocol make sure that ramping times are consistent.
2. Incubator or water bath at 37 °C to perform the overnight incubation when amplifying whole phage genomes.
3. Spectrophotometer to accurately quantify DNA or RNA template. Use quartz cuvettes if quantifying nucleic acid in a spectrophotometer that requires them. A NanoDrop spectrophotometer can be especially useful if one has only limited amounts of template (NanoDrop Technologies, Wilmington, DE; http://www.nanodrop.com/).
4. Gel documentation system or UV transilluminator and camera.

3 Methods

3.1 Template Preparation

A 1 µl sample of a 1:5 dilution of a $\sim 10^{11}$ phage stock diluted in dH_2O may produce a suitable template without having to prepare pure DNA through extraction. An alternative method is to pick a phage plaque and re-suspend it in 50 µl of water. Leave at room temperature for 30–60 min and then boil for a few minutes in a water bath *(13)*. Up to 25 µl of this cleaned phage lysate may be necessary for PCR (*see* **Note 2**). DNA may also be extracted from PFGE experiments by performing two rounds of freezing and thawing of the plugs or bands of interest followed by centrifugation at $1,500 \times g$ to sediment the agarose. The supernatant can be directly used as a template for PCR. This type of approach is often useful when probing community viral DNA, following PFGE analysis to look for the presence of particular genes *(14)*.

If these approaches do not work then a DNA extraction has to be carried out. *See* **Chapters 2, 23 and 34**. Generally, 1–100 ng of high quality DNA template is necessary per reaction. As for any DNA, extracted template should be stored in either ultrapure water or in Tris buffer (pH 7). For reference to appropriate chapters to produce suitable template, please see above.

3.2 Target DNA (Template) for Sequencing

Sequencing of phage genes may be performed from PCR products in the usual way. Alternatively, they may be sequenced directly from phage DNA prepared from Genomiphi (GE Healthcare) or a similar enzyme. This is useful if either it is not possible to get a PCR to work, or it is necessary to walk out from a known sequence to genes that are not known. Refer to manual for instructions. In brief, the template is denatured and then left at 37 °C overnight in the presence of the DNA polymerase from φ29 in the case of Genomiphi and random hexamers. The reaction is denatured by heating at 60 °C for 5 min. The resulting DNA is quantified by absorbance at 260 °C, 1 μg of DNA per sequencing run is required.

3.3 Primer Design

Primers are designed in the usual way and it is a good idea to use software to design them to avoid problems with dimmers, runs and hairpins and appropriate GC contents and melting temperatures. For information on software, see http://molbiol-tools.ca/PCR.htm. The better the primers, the fewer downstream problems there are with PCR. Commonly used primer sequences for diversity screening of marine bacteriophages are given in **Table 5.1**. These give the exact sequences for cyanophage-specific primers from the *Myoviridae (5, 15, 16, 17)* and for more general members of the T4-type phages *(6)*. The primers also illustrate the amount of degeneracy required even in these conserved genes in order to detect as many different isolates as possible.

3.4 Amplification Reaction

There is nothing different about amplification conditions for amplifying genes from bacteriophage DNA as compared with other template. In brief therefore, denature the DNA template for 5 min at 95 °C. Perform 20–30 cycles of denaturation at 95 °C (30 s), annealing at the predicted melting temperature of the primer (30 s) (*see* **Note 3**), and extension at 72 °C for (the time here depends on the product length to be amplified but a general rule of thumb is 30 s per 500 bp). Perform a final elongation at 72 °C for 5 min and store at 4 °C (*see* **Note 4**).

3.5 PCR Enhancers

A number of additives can be included to encourage PCR to work more efficiently. These can be particularly helpful when DNA has been obtained from environmental sources that contain inhibitory substances such as from blood,

Table 5.1
To show commonly used primers in exploring myoviruses diversity

Phage gene	Primer direction and name	Sequence 5′ – 3′	Reference:
Gp20	F CPS1	GTAG(T/A)ATTTTCTACATTGA(C/T)GTTGG	*(3)*
	R CPS2	GGTA(G/A)CCAGAAATC(C/T)TC(C/A)AGCAT	
Gp20	F CPS3	TGGTA(T/C)GT(T/C)GATGG(A/C)AGA	*(17)*
	R CPS4	CAT(A/T)TC(A/T)TCCCA(A/T/C)TCTTC	
	R CPS8	AAATA(C/T)TT(G/A/T)CCAACA(A/T)ATGGA	
Gp20	R G20-2	(G/C)(A/T)(A/G)AAATA(C/T)TTICC (A/G)AC(A/G)(A/T)A(G/T)GGATC	*(5)*
Gp23	Mzia 1	GATATTTGIGGIGTTCAGCCIATGA	*(6)*
	Mzia 2	CGCGGTTGATTTCCAGCATGATTTC	

Sequences from which to design primers are obtained from GenBank. For general sequence recovery information on how to do this. see http://molbiol-tools.ca/GenBank.htm.

faeces, soil or aqueous environments. Useful additives include DMSO (dimethyl sulfoxide), betaine (*N,N,N*-trimethylglycine=(carboxymethyl)trimethylammonium), formamide, non-ionic detergents, such as Triton X-100, Tween 20 or Nonidet P-40 (NP-40), TMAC (tetramethylammonium chloride), 7-deaza-2′-deoxyguanosine (dC^7GTP), BSA (bovine serum albumin), and the T4 gene 32 protein. These inhibitors generally influence the structure of DNA, reducing secondary structure and allowing the template to be more amenable to the PCR. They are all useful in specific circumstances, and if used incorrectly can do more harm than good. Further information can be found in http://www.staff.uni-mainz.de/lieb/additiva.html.

3.6 Detection and Analysis of Reaction Product

Again there is nothing specifically phage related to be done here. Make a 0.8–4% agarose gel in 1X TAE buffer. The percentage of agarose depends on the size of the product. A 500 bp–1 kb fragment will be perfectly visualised in a 1% gel. Weigh out the appropriate amount of agarose and add to the buffer. Melt in the microwave for around 2 min depending on the amount needed. Cool the gel either by swirling under a cold water tap or on the bench. Add 5 μL/50 mL of ethidium bromide and pour into casting tray.

Load around 5 μl of product from the PCR. Run the amplified fragment alongside DNA ladder. Depending on the size of the gel run between 60 and 240 V through it until the dye front has migrated at least 5 cm.

4 Notes

1. The primers usually arrive lyophilised and the amount in nanograms (or nanomoles) synthesised is given in the paper work. Add 1 ml of ultra-pure water to the lyophilised primer and to calculate a working concentration of 10 μM, use the equation $c_1 v_1 = c_2 v_2$, where c is concentration and v is volume. For example, if you have added 1 ml of water to 50 ng of primer, to make the 10 mM working stock in 100 μL you will need to add 20 μL of concentrated primer and 80 μL of ultra-pure water.
2. Adding too much DNA can result in non-specific amplification. When amplifying from bacterial DNA, when screening for phage products of phage-encoded products, more DNA may be required (up to 1000 ng). Care must be taken to limit the amount of Taq polymerase inhibitors as much as possible (such as detergent, EDTA and traces of phenol/chloroform). Therefore, diluting DNA template can result in a successful reaction as it may dilute out an inhibitors present.
3. This can be predicted by in the software package that you design your primers in or a calculator is available from http://www.promega.com/biomath/default.htm, the information will also be included on the details when the primers are delivered.
4. If the PCR is set-up overnight, do not leave on 4 °C until morning as this puts excessive strain on the PCR machine and will significantly shorten its life.

References

1. Casas, V., et al., *Widespread occurrence of phage-encoded exotoxin genes in terrestrial and aquatic environments in Southern California*. Fems Microbiology Letters, 2006. **261**(1): 141–149.
2. Millard, A., et al., *Genetic organization of the psbAD region in phages infecting marine Synechococcus strains*. Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(30): 11007–11012.
3. Fuller, N.J., et al., *Occurrence of a sequence in marine cyanophages similar to that of T4 g20 and its application to PCR-based detection and quantification techniques*. Applied and Environmental Microbiology, 1998. **64**(6): 2051–2060.
4. Marston, M.F. and J.L. Sallee, *Genetic diversity and temporal variation in the cyanophage community infecting marine Synechococcus species in Rhode Island's coastal waters*. Appl Environ Microbiol, 2003. **69**(8): 4639–47.
5. Short, C.M. and C.A. Suttle, *Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments*. Applied and Environmental Microbiology, 2005. **71**(1): 480–486.
6. Filee, J., et al., *Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere*. PNAS, 2005. **102**(35): 12471–12476.
7. Petrov, V.M., et al., *Plasticity of the Gene Functions for DNA Replication in the T4-like Phages*. Journal of Molecular Biology, 2006. **361**(1): 46.
8. Miklic, A. and I. Rogelj, *Characterization of lactococcal bacteriophages isolated from Slovenian dairies*. International Journal Of Food Science and Technology, 2003. **38**: 305–311.

9. Deveau, H., et al., *Biodiversity and Classification of Lactococcal Phages*. Appl. Environ. Microbiol., 2006. **72**(6): 4338–4346.
10. Balding, C., et al., *Diversity of phage integrases in Enterobacteriaceae: development of markers for environmental analysis of temperate phages*. Environmental Microbiology, 2005. **7**(10): 1558–1567.
11. Sullivan, M.B., et al., *Prevalence and Evolution of Core Photosystem II Genes in Marine Cyanobacterial Viruses and Their Hosts*. PLOS, 2006. **4**(4): 1344–1357.
12. Baker, A.C., et al., *Identification of a Diagnostic Marker To Detect Freshwater Cyanophages of Filamentous Cyanobacteria*. Appl. Environ. Microbiol., 2006. **72**(9): 5713–5719.
13. Kutter, E. and A. Sulakvelidze, *Bacteriophages: Biology and Applications*. 1 ed. 2005, Boca Raton: CRC Press. 510.
14. Sandaa, R.-A. and A. Larsen, *Seasonal Variations in Virus-Host Populations in Norwegian Coastal Waters: Focusing on the Cyanophage Community Infecting Marine Synechococcus spp*. Appl. Environ. Microbiol., 2006. **72**(7): 4610–4618.
15. Fuller, N.J., et al., *Clade-specific 16S ribosomal DNA oligonucleotides reveal the predominance of a single marine Synechococcus clade throughout a stratified water column in the Red Sea*. Applied and Environmental Microbiology, 2003. **69**(5): 2430–2443.
16. Wilson, W.H., et al., *Analysis of cyanophage diversity and population structure in a south-north transect of the Atlantic ocean*. Bulletin de l'Institut océanographique, Monaco, 1999. **19**: 209–216.
17. Zhong, Y., et al., *Phylogenetic diversity of marine cyanophage isolates and natural virus communities as revealed by sequences of viral capsid assembly protein gene g20*. Applied and Environmental Microbiology, 2002. **68**(4): 1576–1584.

Chapter 6

In Silico Identification of Genes in Bacteriophage DNA

Andrew M. Kropinski, Mark Borodovsky, Tim J. Carver, Ana M. Cerdeño-Tárraga, Aaron Darling, Alexandre Lomsadze, Padmanabhan Mahadevan, Paul Stothard, Donald Seto, Gary Van Domselaar and David S. Wishart

Abstract

One of the most satisfying aspects of a genome sequencing project is the identification of the genes contained within it.These are of two types: those which encode tRNAs and those which produce proteins. After a general introduction on the properties of protein-encoding genes and the utility of the Basic Local Alignment Search Tool (BLASTX) to identify genes through homologs, a variety of tools are discussed by their creators. These include for genome annotation: GeneMark, Artemis, and BASys; and, for genome comparisons: Artemis Comparison Tool (ACT), Mauve, CoreGenes, and GeneOrder.

Key words: tRNA, transfer RNA, CDS, ORF, Software, Internet, ACT, Artemis Comparison Tool, Artemis, BASys, Blast, CoreGenes, Dotplot, frameshift, GeneMark, GeneOrder, intron, Mauve, Online Analysis Tools.

1 Introduction

One of the most satisfying events of working with a phage occurs when after several years of studying the physiology and genetics, one obtains its genome sequence. Many of the hypotheses that may have been generated now can be put into perspective of the entire genome sequence rather than the sequence of limited regions. Further hypotheses can be generated as a result of the identification of viral genes through homology and motif analyses. Conversely, it can also be extremely frustrating if homologs do not exist and when the translation initiation positions prove difficult to define.

Martha R. J. Clokie, Andrew M. Kropinski (eds.), *Bacteriophages: Methods and Protocols, Volume 2: Molecular and Applied Aspects, vol. 502*, © 2009 Humana Press, a part of Springer Science+Business Media
DOI 10.1007/978-1-60327-565-1_6 Springerprotocols.com

Bacteria and their viruses encode two major types of genes—those that specify proteins and those that encode structural RNAs. These algorithms, which identify the former, will not locate the later. In the following sections, we will deal with the identification tools and the logic used in identifying genes. In addition, tools for the further characterization of the gene products will be discussed. The emphasis will be on free Internet resources and platform-independent software, such as Java or Perl.

The DNA sequence itself provides data for the determination of: the mol% G+C, the variation of base composition along with the length of the DNA (which provides information concerning possible recombinational events that contributed to the evolution of the phage), DNA skew analysis which often provides useful information concerning the replication origin *(1)* and 'of course' the location of genes (see below), promoters and terminators (*see* **Chapter 28**). Tools for the analysis of base composition, repeats and restriction endonuclease cleavage sites can be found in Online Analysis Tools, specifically http://molbiol-tools.ca/DNA_composition.htm, http://molbiol-tools.ca/Repeats_secondary_structure_Tm.htm, and http://molbiol-tools.ca/Restriction_endonuclease.htm.

2 Transfer Ribonucleic Acids (tRNAs)

While genes encoding tRNAs have been observed in all branches of the *Caudovirales*, they are most commonly found in the members of the *Myoviridae* with large genomes and only occasionally in members of the *Podoviridae (2,3)*. These genes have been identified, singly or in multiple copies, in phages infecting cyanobacteria *(4,5), Aeromonas (6), Pseudomonas (7), Vibrio (8)*, coliforms *(9, 10, 11, 12), Salmonella (13), Streptomyces (14, 15), Mycobacterium (16, 17)*, and *Listeria (18)*. The best tools for locating these elements in DNA sequences is tRNAscan *(19)* which can be accessed at http://lowelab.ucsu.edu/tRNAscan-SE/or ARAGORN (http://130.235.46.10/ARAGORN/) *(20)*. These programs provides information on the location, type of tRNA molecule, and anticodon used. They also provides a diagram of the tRNA species identified in cloverleaf format. One might want to follow-up with a nucleotide search using the Basic Local Alignment Search Tool (BLASTN) at the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/BLAST/) against the non-redundant (nr) databases to identify potential prophage and phage homologs.

3 Protein-Encoding Genes

A predicted segment of DNA which might be translated into a polypeptide can be called an ORF (*O*pen *R*eading *F*rame) or CDS (*CoD*ing *S*equence). The following are the common traits of functional genes located within regions bounded by two stop codons (TAG, TGA, and TAA):

1. *Numbers of nucleotides*: Most of the gene identifying software packages limit the size of identifiable genes to greater than 100—150 nucleotides, but smaller genes have been recognized in phages. For example, enterobacterial phage Sf6 gene Sf6p45 encodes a 27 amino acid NinA homolog, while the inhibitor of the alpha-lipopolysaccharide polymerase in *Pseudomonas* phage D3 is 31 amino acids *(21)*. These small genes would be found if the software search parameters were set to a minimum of 75 nt. However, this setting is likely to generate many false positives. At the other end of the spectrum, huge genes have been identified in cyanophages. For example, gene PSSM4_0080 encodes a protein containing 7312 amino acid residues *(22)*.
2. *Start codons*: The usual start codons are AUG (ATG in the DNA sequence), GUG (GTG), and UUG (TTG) which account for over 98% of initiation codon usage in prokaryotes. In an analysis of 620 complete prokaryotic genome sequences, Villegas and Kropinski showed that these three codons are used, on an average, 80.1% [AUG], 11.6% [GUG], and 7.8% [UUG] *(23)*. It is noteworthy that the use of the GUG initiation codon increases, and AUG usage decreases with the GC content of the genome. It is, therefore, very important to use gene finding software which recognizes alternative start codons.
3. *Stop codons*: The translational stop codons are: (TAA [UAA, ochre], TGA [UGA, opal or umber], and TAG [UAG, amber).
4. *Ribosome-binding site*: Upstream (5′) of coding sequences one will frequently observe a ribosome-binding (RBS) or Shine-Dalgarno *(24, 25)* site approximately 3–15 nt upstream to the gene [most frequently 4–9 nt]. This sequence usually resembles, in the Enterobacteriacae, a subset of TAAGGAGGT (UAAGGAGGU) or its part. Some exceptions exist, such as the lambdoid phage repressor genes which frequently lack external RBSs *(26)*.
5. *Proximity*: With rare exceptions, phages genes densely packed in their genomes contain many overlaps between the adjacent genes. In rare cases, one ORF may be initiated well within another gene which would result, if the ORF start is situated in the same frame, in proteins with a common C-terminus or,

if the start is situated in another reading frame, in completely different proteins. The latter are more common in the small phages, such as φX174 than in members of the *Caudovirales*.

6. *Codon usage bias*: This is an important feature of the protein-coding region compositions which have been used to delineate prokaryotic genes. However, most phages have undergone extensive horizontal gene transfer *(27, 28)* which can make gene identification difficult by the species-specific tools employing the parameters of the host codon bias.

3.1 Oddities in Phage DNA Sequences

Two rare phenomena can influence automatic gene annotation. These are the presence of introns and translational frameshifting. While relatively rare group I self-splicing introns have been discovered in a variety of phages including T4 and its relatives *(9, 29, 30)*, as well as coliphages K1F *(31)* and W31 *(32)*, streptococcal phages *(33, 34)*, *Lactococcus* phages Tuc2009 *(35)* and r1t *(36)*, *Bacillus* phages Bastille and SPβ *(37, 38)*, *Synechococcus* phage S-PM2 *(39)*, and *Staphylococcus* phages K and Twort *(30, 40)*. The suggestions that introns maybe present is a specific gene initially comes from BLASTX analysis of the DNA sequence. In the following example (**Fig. 6.1**), a sequence encoding a DNA polymerase gene from bacteriophage K1F was analyzed. The latter reveal that the regions of homology are separated by an ORF-bearing sequence similarity to a homing (HNH) endonuclease.

Frameshifting (or ribosomal slippage) is observed when the translational machinery slips forward or backwards on the message (mRNA) and continues to translate in another reading frame. This event is associated with a "slippery sequence" but since such sequences can vary significantly it makes predictions based solely on sequence data very difficult. Usually the existence of ribosomal slippage is suggested by the proteomic data. This was the case with the structural proteins of enterobacterial phages T7 *(41)*, λ *(42)*, P2 *(43)*, and MB78 *(44)* and *Listeria* phage PSA *(18)*. In each case, two *cis*-acting sequences were involved, the slippery site and a downstream pseudoknot structure *(45)*. The latter structure can be identified using pknotsRG at http://bibiserv.techfak.uni-bielefeld.de/pknotsrg/ *(46)*. In the case of alkaliphilic phage

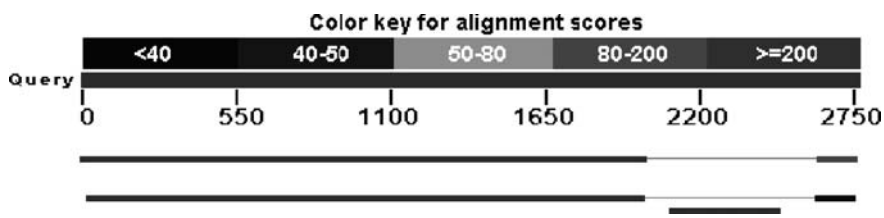


Fig. 6.1. BLASTX analysis of the coliphage K1F DNA sequence (NC_007456) from nucleotides 12921–14936. The regions of homology (*thicker lines*) are separated by a thin line, and the latter region of the sequence contains a homing endonuclease.

BCJA1c, the sequence of the integrase gene was found to have a persistent "error"—an in-frame stop codon overlapping a slippery site (AAAAAAAG). This combination forms a novel method for regulating integration *(47)*.

4 Gene Analysis

A wide variety of online tools are available for detection of genes in prokaryotes and their viruses. These include, in alphabetical order: Artemis, BASys, BLAST, FSGENESB, GeneHacker, GeneMark, Generation, and Glimmer. Because of its unique abilities, BASys will be discussed in the next section.

4.1 BLASTX

As such BLAST *(48)* does not identify open reading frames, merely regions which encode proteins with relatedness to existing proteins in the database. It is of limited utility with unique viruses, such as some of the archael phages whose genes lack homologs. But for an increasing number of phages, it is a highly useful tool both in the initial and in the final stages of an annotation project. The following are our recommendations on how to get the most out of it.

Divide your genomic sequence into a series of approximately 10 kb overlapping fragments, i.e., 0–10 kb, 9.5–20 kb, 19.5–30 kb, etc. Use the translated BLAST (BLASTX) search engine at NCBI (http://www.ncbi.nlm.nih.gov/BLAST/) setting the "Format" option to "Descriptions = 1000" and "Alignments = 1000." This wide window will increase the probability of useful hits if a highly represented gene, such as an integrase, is present in your sequence. If you wish you can adjust the "Options" from the default "nr" setting to "VIRUSES [ORGN]" to only scan for viral homologs. Please note, that as of writing this chapter, there is no database at NCBI specifically for prophage sequences. Once the search results appear, save the page to an appropriately labeled directory on your computer. We have noted that Mozilla Firefox (http://www.mozilla.org/products/firefox/) works better than Microsoft Internet Explorer at saving the entire page in html format which can subsequently be opened.

Homologs are represented as a truncated set of colored bars across the screen (*see* **Fig. 6.2**).

Click the topmost bands and you will move in the page to the homolog and its alignment with your DNA sequence translated in one of the six reading frames. This information will give you about:

1. The existence of homologs and their location on the DNA sequence.

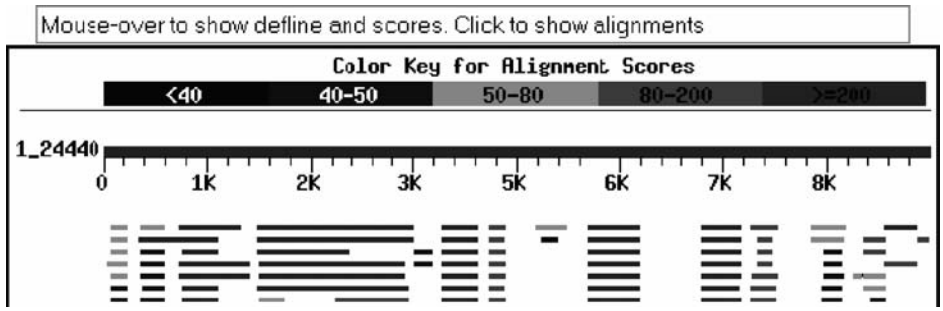


Fig. 6.2. A portion of the BLASTX showing the results obtained for a 9 kb region of phage ε34. The regions of the sequence which when translated result in significant "hits" are illustrated as a series of differentially *colored bars*. Please note that the lack of homologs between 6.2 and 6.7 kb does not indicate that this region lacks genes.

2. The existence of pseudogenes in the sequence. These are relatively common in prophage sequences and are identified as regions of homology which do not correspond to an obvious gene.
3. It will answer questions concerning direction such as is the gene on the plus (top) or minus (bottom) strand and in what reading frame.
4. It may lead to the identification of possible errors in the sequence. These can be of two types: in-frame stop codons (* in **Fig. 6.3**A) and situations where homologs are out-of-frame such that the relationship to a single homolog is to be found on two (or more) reading frames (**Fig. 6.3B**). If these are observed the region should be resequenced.

4.2 Artemis

Artemis is a nucleotide sequence viewer and annotation tool (**Fig. 6.4**) *(49)*. It is freely available and can be downloaded from the Sanger Institute web page <http://www.sanger.ac.uk/software/Artemis> where it is developed and maintained. It is written in Java making it plat-

```
gi|8439603|gb|AAF75025.1|     antirepressor [Enterobacteria phage P22]

          Length = 61

 Score =  120 bits (300), Expect = 1e-26
 Identities = 60/61 (98%), Positives = 60/61 (98%)
 Frame = +1

Query: 1   MYKKDVIDHFGTQRAVAKA*GISDAAVSQWKEVIPEKDAYRLEIVTAGALKYQENAYRQA 180
           MYKKDVIDHFGTQRAVAKA GISDAAVSQWKEVIPEKDAYRLEIVTAGALKYQENAYRQA
Sbjct: 1   MYKKDVIDHFGTQRAVAKALGISDAAVSQWKEVIPEKDAYRLEIVTAGALKYQENAYRQA 60

Query: 181 A 133
           A
Sbjct: 61  A 61
```

Fig. 6.3A. Alignment in the BLASTX output showing an inframe stop codon illustrated as an asterisk (*) in the P22 antirepressor gene.

```
☐gi|11611120|emb|CAC18561.1|   putative 0.45 protein [Bacteriophage phiYeO3-12]

          Length = 66

 Score = 75.1 bits (183), Expect(2) = 4e-29
 Identities = 35/36 (97%), Positives = 36/36 (100%)
 Frame = +1

Query: 1   MSKLLATSKIEGQCTVTLREYYHGSMGSTYVVRYGQ 108
           MSKLLATSKIEGQCTVTLREYYHGSMGSTYVVRYG+
Sbjct: 1   MSKLLATSKIEGQCTVTLREYYHGSMGSTYVVRYGK 36

 Score = 74.3 bits (181), Expect(2) = 4e-29
 Identities = 31/31 (100%), Positives = 31/31 (100%)
 Frame = +2

Query: 107 KQVTHWVNPILAQEDYQSCVLHQTTCAGWND 199
           KQVTHWVNPILAQEDYQSCVLHQTTCAGWND
Sbjct: 36  KQVTHWVNPILAQEDYQSCVLHQTTCAGWND 66
```

Fig. 6.3B. An out-of-frame mutation in gene 0.45 of phage φYeO3-12 as shown by a change in reading frame of the homologous sequence.

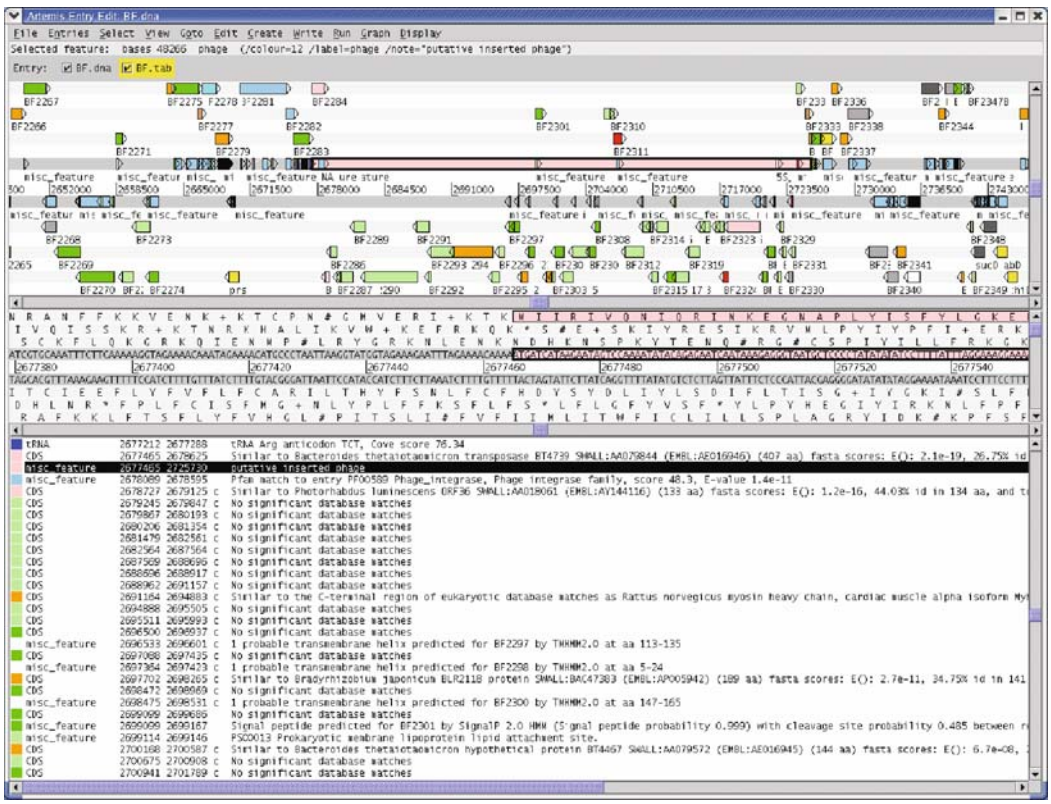


Fig. 6.4. This is the Artemis interface with part of the *Bacteroides fragilis (88)* NCTC9343 genome loaded in. It shows a putative inserted phage (*long highlighted pink box*) with the coding sequences (CDSs) annotated in and around it. At the top of the interfaces, the main menus can be seen. The following row shows information about the selected feature. In the next row, down are the names of the files read in. Beneath that are the main display views, showing a zoomed out and zoomed in representation of the sequence. The bottom window shows a scrollable list of features.

form independent and it can be run on UNIX, Windows, and MacOSX.

There are different levels of granularity at which sequences can be viewed in Artemis. This ranges from the base and amino acid level up to the complete genome. When Artemis is first launched, it provides two main views of the sequence. Both display the sequence with the forward and reverse strands top and bottom along with the associated six-frame translation. The top sequence display provides an overview of a larger region ($\sim$ 10 kb), whereas the display underneath shows a smaller region ($\sim$0. 1 kb) and actually displays the amino acids and nucleotides. The zoom level can be adjusted with the slider at the side of the window, such that it is possible to zoom out completely to view the entire genome.

Common formats such as EMBL GenBank, fasta, and GFF can be used to load the sequences and features into Artemis. These can be opened either by giving the file name as a command line argument or from the Artemis "File" menu. Alternatively, a web or ftp address can be used to download and open the sequence. Multiple sequences can be loaded from a multiple fasta file and these are shown appended to each other in the Artemis interface, with the separate sequences identified as features.

The features loaded into Artemis can be selected and their annotation displayed and edited. The sequence and features can be navigated in various different ways. The scroll bars beneath the sequence can be used to move up and down the sequence. Also under the "GoTo" menu, there is a "Navigator" tool. This can be used to define a base number, gene name, base pattern or amino acid pattern to search for. There are a number of other options that make this tool a very flexible method of locating regions of interest.

Another Artemis tool for moving about the genome and identifying the regions is the "Feature Selector." Again there are a number of different ways this can be used. For example, it is possible to select and list features with a particular key or qualifier containing a given text, such as phage, prophage, and/or bacteriophage, depending on the annotation associated with the feature(s).

In Artemis, features can be created in a number of different ways. Multiple sequence and feature files, known as entries, can be read in and overlaid on top of each other. So, for example, the output from a gene prediction calculation can be read in over the sequence. Entries can be viewed either on the six-frame translation or on the separate lines by selecting the "One Line Per Entry" option. This makes it a powerful tool for build up data about predicted features on the genome.

Alternatively, features can be automatically generated, for example, by creating CDS features in the open reading frames (defined as the regions between stop codons). The minimum

length of the open reading frames can be defined (the default being 100 bases). The features created can also be trimmed to any of the three common different bacterial start codons or manually adjusted by dragging the ends. It should be noted that not all open reading frames are coding, and this is not recommended as a method of gene prediction for any but the shortest sequences.

Features calculated from external analyses, such as gene prediction programs, can be loaded in as described above. However, there are additional tools in Artemis to assist in locating genes. For example, codon usage tables can be loaded to generate codon usage plots, and frame-specific G+C content plots can be displayed. The graphs are displayed above the genome sequence, and the results of many other algorithms (internal and external) can be shown in this way, and scrolled and zoomed along with the sequence, in order to assist in gene identification and verification. In addition, percentage G+C/A+G graphs can be displayed and used to identify areas of low or high G+C content, such as prophages inserted in a given genome (*see* **Fig. 6.5**).

Optional tools can be installed and integrated via shell scripts. Typically, these include BLAST and fasta searches, which can be

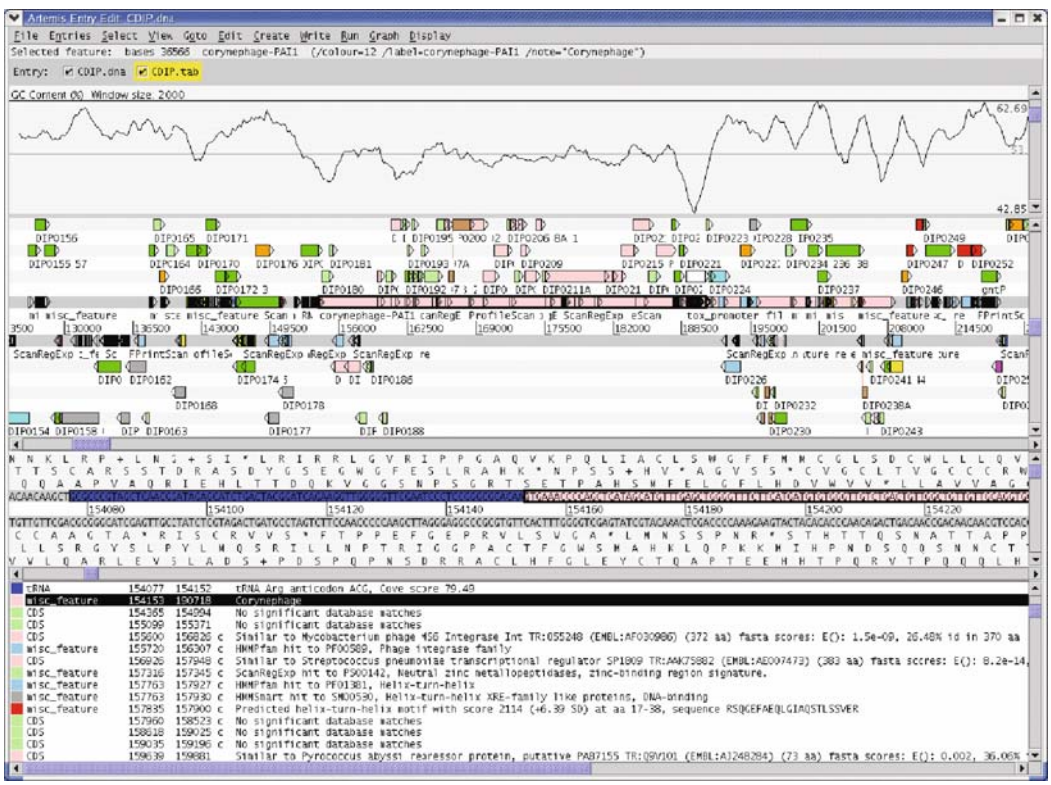


Fig. 6.5. Artemis window displaying the corynephage in the *Corynebacterium diphtheriae* NCTC13129 genome *(89)* along with a G+C content (%) graph calculated with a 2 kb sliding window. Note the drop in the G+C content in the region highlighted corresponding to the corynephage. Also shown is the diphtheria toxin precursor gene, white box tox on the right-hand end of the corynephage within its own region of lower G+C

.

set up so that they can be called from Artemis. This involves editing an options file in the distribution (e.g., etc./options) to add the databases (e.g., UniProt) and customizing the run scripts (e.g., "etc./run_blastp") for the computer environment they are being run on. This enables the user to select a feature or multiple features and run these analyses on them. When the analyses have finished, the results are viewed in a window in Artemis or can be sent to a web browser.

4.3 Gene Finding in Phage Genomes Using the GeneMark Family Algorithms

Gene finding algorithms of GeneMark family *(50)* were developed for analysis of DNA sequences from various sources: prokaryotes, eukaryotes, EST/cDNA, viruses, and phages. We focus here on gene finding in phages. There are two aspects of computational gene finding that should be considered separately: prediction algorithm and algorithm for parameter estimation.

Organization of protein coding genes in phage genomes does not differ significantly from organization of genes in genomes of their prokaryotic hosts. Therefore, methods developed for gene finding in prokaryotes, GeneMark *(51)* and GeneMark.hmm *(52, 53)*, can be applied to phage genomes. Both algorithms use the formal concepts of genes and intergenic regions, initiation and termination codons, as well as other relevant sites, such as RBS (ribosomal-binding sites). In the algorithms, implementation different genetic codes can be used with amino acid translation table modifiable by the algorithm parameters. Genes can be predicted on both strands of double helix and various types of gene overlaps are allowed. Sequencing errors can be identified by the GeneMark algorithm. Several rare aspects of gene organizations do not have support in phage algorithms, such as: programmed frame shifts, introns, and complete gene overlaps.

The GeneMark algorithm is the implementation of the Bayesian pattern recognition approach. It identifies the functional meaning of the sequence using the Bayesian posterior probability that the local statistical measures of a nucleotide sequence fit to the three-periodic Markov model of protein-coding region. The GeneMark.hmm algorithm is using similar technique "on top" of the hidden Markov model, describing transitions between protein-coding and non-coding regions. This algorithm finds the most probable parse of the whole sequence into genes and intergenic regions. GeneMark has additional graphical output that provides convenient view of the prediction function for a given sequence (**Fig. 6.6**). GeneMark is not restricted by the HMM grammar as GeneMark.hmm. Therefore, GeneMark is more robust and less affected by deviations from the ideal grammar (such as sequencing errors). On the other hand, GeneMark.hmm has better accuracy in prediction of short genes and in finding protein translation initiation sites. Thus, these two algorithms have complementary properties.

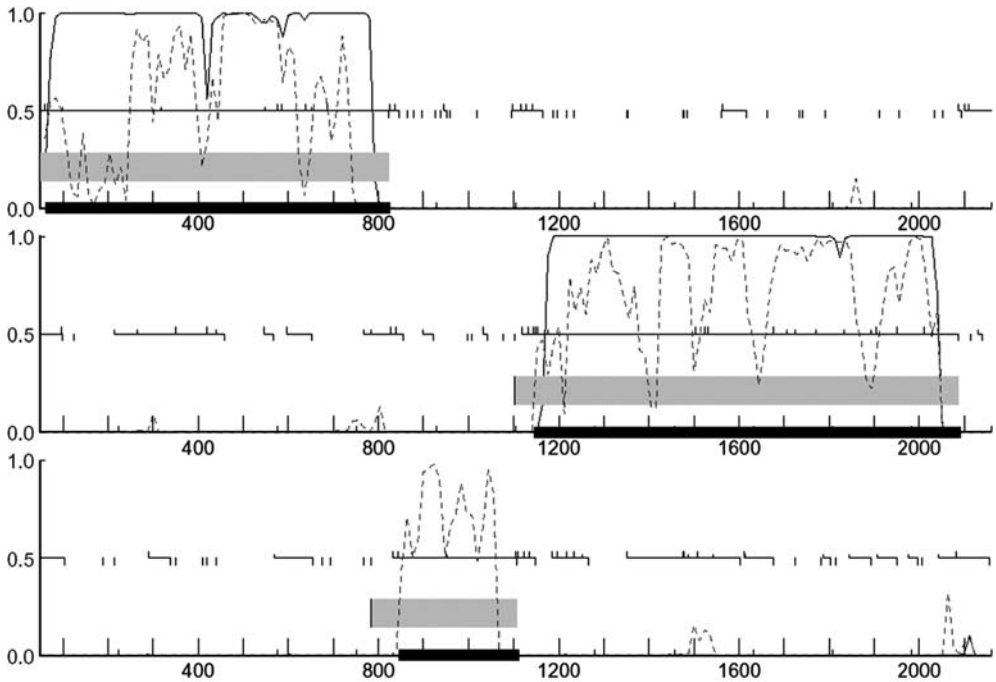


Fig. 6.6. Graphical output of GeneMark and GeneMark.hmm. *The solid black and dashed lines* indicate the coding potential calculated by the GeneMark program using the GeneMarkS and Heuristic parameters, respectively. *The thick black horizontal bars* indicate the predictions made by GeneMark.hmm.

Statistical models necessary for gene finding are estimated (trained) by the algorithms which are not the same for genomes of phages and their hosts. Phage genomes are significantly smaller than prokaryotic genomes. Therefore, training methods developed for large genomes cannot immediately be applied to phage genomes, as some parameters may be determined with prohibitively large errors. The problem of training for gene prediction in viruses and phages was addressed in VIOLIN paper *(54)*. The suggested training procedure makes use of sequence composition for the Heuristic parameter estimation *(55)* if the sequence is short. Otherwise, if the DNA sequence is long enough, the model parameters are determined by the unsupervised parameter estimation method GeneMarkS *(53)*.

The Heuristic algorithm estimates codon usage frequencies from the genomic sequence composition. It is using the correlations patterns observed in several complete prokaryotic genomes. The GeneMarkS algorithm iteratively derives the parameters of the models from a large anonymous genomic sequence via alternating prediction and parameter estimation steps. This algorithm identifies frequency patterns typical for protein coding/intergenic sequences as well as for short motifs, such RBS. Model parameters derived with either of the training methods can be used with

either GeneMark or GeneMark.hmm. Gene prediction made with GeneMarkS parameters usually show high accuracy in localization of exact boundaries of genes (translation initiation starts). They also show high specificity in predicting genes. Predictions based on Heuristic models exhibit high sensitivity. The Heuristic model is capable to predict genes with unusual codon usage (frequently related to lateral gene transfer between genomes). Predictions made with the Heuristic and the GeneMarkS parameters are combined inside the algorithm of GeneMark.hmm *(53)*. The GeneMark program uses these models one at a time.

Since phages use the translation machinery of their hosts parameters for gene finding can be at times inferred from host genomes. However, some phages can encode their own tRNAs and can have significantly different codon usage from the host, which limits application of this method.

The flowchart of the model parameter estimation for gene prediction in phage genomes is shown in **Fig. 6.7**.

Below we describe a pipeline which includes some rules and thresholds, practically useful for the analysis of phage genomes *(54)*. This pipeline demonstrates stable performance in automatic mode, but for some genomes additional manual intervention and parameters adjustment can further improve annotation quality *(56)*. Many steps described in the pipeline can be performed using the web service at http://opal.biology.gatech.edu/GeneMark/ *(50)* (marked as "web" in pipeline description). For some steps, the local version of the program and some additional scripts are required (marked as "local"; note that all the "web" steps can be done in the "local" mode).

The phage genome annotation pipeline:

1. Mask tRNA genes, especially in genomes with low G + C content (not as important in high G + C genomes). [local]
2. Apply GeneMark.hmm with heuristic parameters. [web]
3. If number of genes predicted is small (less than 50), go to the refinement of the translation initiation sites [step 4], else enter GeneMarkS [step 6].
4. Heuristic parameters have limited use for the start codon prediction. If the host genomic data are available, sequence of host can be analyzed, the RBS model derived and use in the algorithm in combination with heuristic model to refine gene start positions [local]. For phage genomes with low G + C, extension of start to longest ORF is appropriate in the absence of the RBS model, as the error introduced by elongation is smaller than error of initial prediction [local]. For phages with high G + C (more than 60%) extension is not recommended.
5. Combine the modified predictions by GeneMark.hmm [step 4] with heuristic predictions by GeneMark [web]. The example of visualization of the GeneMark prediction is shown in

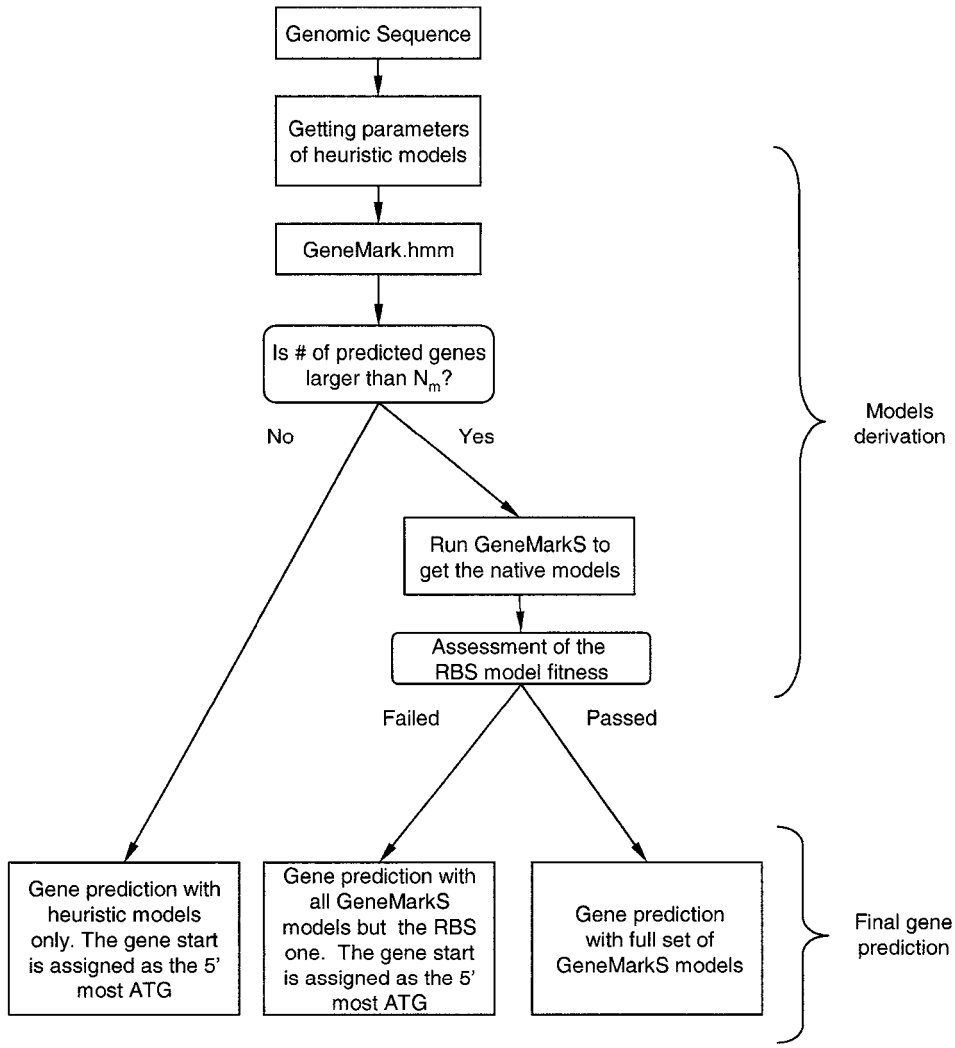


Fig. 6.7. Flowchart of the statistical gene identification procedure applied to a complete genome of a virus of a prokaryotic host.

Fig. 6.1. Frame shifts/sequencing error can be detected by GeneMark (in the text output or in the graph). Visual analysis of graph can help in finding unusual gene organization in the phage genome. Here comes the end of the analysis.

[Continuation from step 3]

6. Run GeneMarkS for the large sequence. If GeneMarkS successfully completes parameters derivation—go to [step 7], if not—move back to [step 4].
7. Check the quality of the derived RBS parameters: whether motif matches to biologically meaningful sequence and if it is narrow localized, with respect to the start codon [local]. If parameters are of reasonable quality, move to [step 8], if not,

initiation site predictions are modified, as described in [step 4], and result of modification is forwarded to [step 8].

8. Run sequence analysis by GeneMark and GeneMark.hmm with combination of Heuristic, and GeneMarkS models [web]. Here comes the end of the analysis.

Note that the automated part of the described procedure was earlier applied to analyze a large collection of virus and phage genomes. Results of prediction are accessible through VIOLIN database at http://opal.biology.gatech.edu/GeneMark/VIOLIN/. Analysis of murine cytomegalovirus virus in *(56)* is an interesting example of the VIOLIN procedure modification.

5 Automatic Annotation Tools

5.1 BASys

The sequencing of complete phage genomes is now fairly routine, while the sequencing of entire bacterial genomes is almost becoming commonplace. Indeed, the ease with which we can generate sequence data now exceeds our capacity to annotate it. This sequence-annotation dilemma has led to the development of automated genome annotation systems. These large and complex programs not only identify the genes, but automatically predict the function, name, and general properties of the gene products. One example of an automated genome annotation system is BASys (Bacterial Annotation System). BASys is a web server that was specifically developed to annotate bacterial and phage genomes *(57)*. It uses more than 30 programs to provide approximately 60 annotation fields for each gene, including gene/protein name, gene/protein function, possible paralogs, and orthologs, molecular weight, isoelectric point, secondary structure, and 3D structure. The textual information returned by BASys is hyperlinked to a navigable graphical map of the genome which provides a convenient interface for genome exploration. Alternatively, BASys results can be searched using server-side text and BLAST searching tools.

The BASys server (http://wishart.biology.ualberta.ca/basys/) provides both anonymous and login-based access for submitting, monitoring, and retrieving genome annotations. When an anonymous submission is made, a secure URL is emailed to the user. This URL provides a progress monitor while the genome is analyzed, and links to the results once they are available. For login access, users must first register with BASys to obtain a password and user ID, which can subsequently be used to submit and monitor multiple genomes.

Genome information is submitted to BASys for annotation using a web-based form. Raw DNA or protein sequence data must be uploaded as a FASTA-formatted file. Also required are the

chromosome topology (circular or linear), the source organism type (phage, Gram-positive bacteria, or Gram-negative bacteria), and a sequence name or ID for monitoring the annotation progress and for identifying the results. If given only genomic DNA sequence, BASys performs its gene predictions using Glimmer *(58)*. Users may select one or more tools and may have consensus gene predictions automatically generated. Alternatively, gene positions can be supplied in a simple tab-delimited format, as an NCBI ".ffn"-formatted FASTA file. Descriptions of the formats and links to examples of the various input files are provided on the submission form.

Once a sequence file has been provided, the BASys annotation engine performs a combination of database comparisons and computational sequence analyses. Translated coding sequences are first compared to several extensively annotated reference databases, including UniProt *(60)*. BASys associates a similarity threshold with each type of annotation in the reference databases. The BLAST similarity score between the query sequences and the database sequences is compared to the thresholds, and qualifying annotations are transitively applied to the query sequence. Several sequence analysis modules are then used to further annotate the sequences: protein family classification is performed using Pfam *(61)*; sequence motifs are identified using PROSITE *(62)*; PredictSPTM (J. Cruz, unpublished data) is used to predict signal peptides and transmembrane domains; and secondary structure is predicted using PSIPRED *(63)*. Some annotations, such as protein molecular weight and isoelectric point, are calculated directly from the query sequences themselves.

The completed genome can be viewed on the BASys server, or downloaded using the links supplied on the results page. BASys provides results in the form of a genome map that is used to browse and evaluate the gene annotations (**Fig. 6.8**).

These maps may be linear (looking like a horseshoe) or circular. The map is implemented using a collection of standard PNG image files and web pages, making it compatible with all current web browsers. Clicking on a gene label opens the corresponding "gene card," which contains the annotations pertaining to the gene in tabular format (**Fig. 6.9**).

Gene card fields derived from external sources are hyperlinked to those sources so that the complete external database records can be viewed. Each gene card is also linked to an "evidence card" containing more detailed descriptions of the source and quality of the annotations. The annotations can also be explored using the text and BLAST searching tools provided on the BASys server. Clicking on a hit returned from one of the search tools opens the relevant gene card and map image.

BASys is continually being upgraded as new databases and analysis modules become available. For the latest informa-

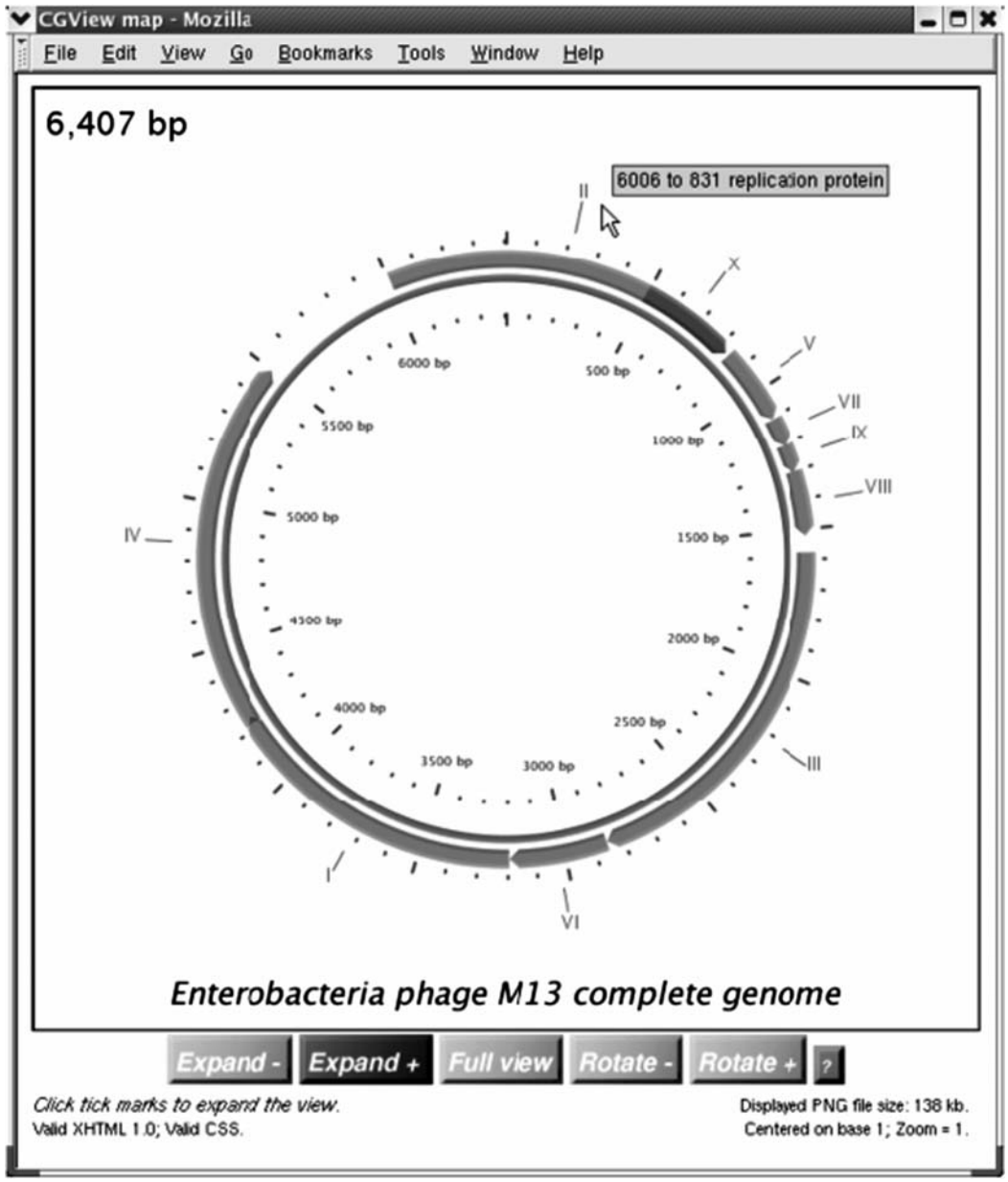


Fig. 6.8. A graphical map generated by BASys using the sequence of the M13 bacteriophage as input. Genes are shown as colored arcs and are labeled. Each label is hyperlinked to the complete set of annotations for the gene. Navigation buttons below the map allow image zooming and rotation. A region of interest can also be expanded by clicking on an adjacent tick mark.

tion on the programs and databases comprising BASys, and for a guide on using BASys, see the documentation available at http://wishart.biology.ualberta.ca/basys/cgi/documentation.pl.

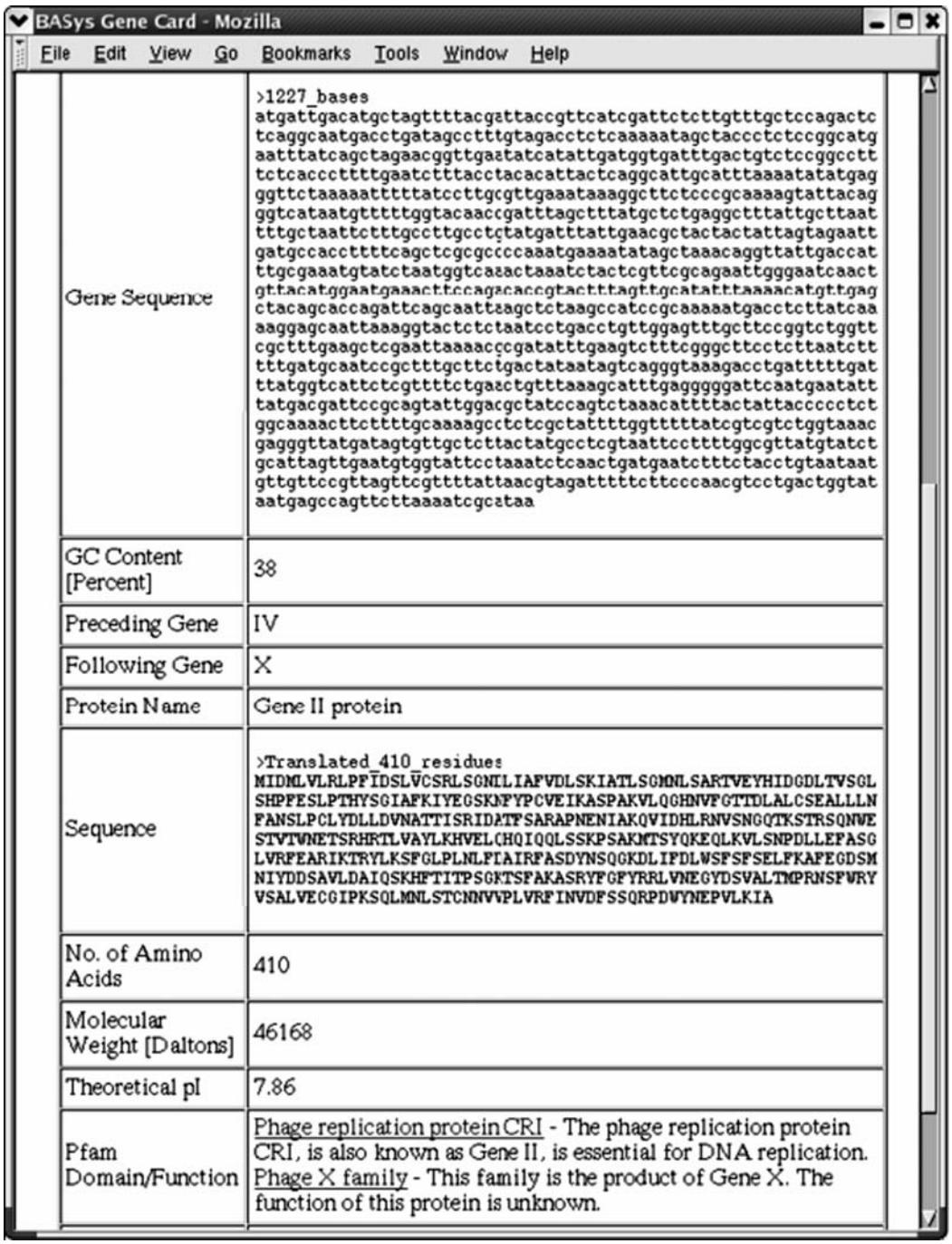


Fig. 6.9. A textual "gene card" generated by BASys. Nearly 60 fields of information are provided for each gene annotated by BASys. Annotations derived from external databases (Pfam, for example) are hyperlinked to the external records for convenient viewing.

6 Comparison of Predictive Tools

The sequence of *Salmonella enterica* serovar Anatum (15 +) bacteriophage ε34 was independently annotated by Sherwood Casjens (University of Utah), Bob Villafane (Ponce School of Medicine), and Andrew Kropinski, and our results were compared with those obtained using the following online tools: AMIGene *(64)*, EasyGene *(59)*, Softberry, Inc. (Mount Kisco, NY) program FGENESB, GeneHacker *(65)*, GeneMark *(50)*, and Generation and Glimmer *(66)* at the ORNL Genome Analysis Pipeline (Oak Ridge National Laboratory, Computational Biology, Oak Ridge, TN, USA). The latter program was also used at PathoGene *(67)*. These programs were used with their default settings, unless they offered the ability to model the predictions based on specific bacteria, in which case *Salmonella* was the preferred option. In each case we looked for the total number of ORFs correctly identified, the number which overlapped with one of ours and the number which were missing or misidentified (**Table 6.1**).

7 Post-Gene Identification Protocols and Tools

When all the genes have been identified, each protein should be subjected to a protein BLAST (BLASTP) search, and in the case of weak homologs to iterative Psi-BLAST analysis *(68)*. In addition to providing information on possible homologs, these tools also provide one with data on potential conserved motifs within the protein sequences. Specifically, BLASTP at NCBI links to the CDD [Conserved Domain Database] *(69)*, pfam [Protein Families] *(61)*, SMART *(70)*, and COGs [Clusters of Orthologous Genes] *(71)*. Another excellent resource which incorporates many of these tools is InterProScan [http://www.ebi.ac.uk/InterProScan/] *(72,73)*. In each of these cases, one wants to look for at least 90% conservation of the motif within the protein under study. Where the motif is less conserved, one should interpret the results with caution. In addition, it is highly recommended that one also submits the protein to programs specifically designed to determine membrane proteins such as TMHMM [http://www.cbs.dtu.dk/services/TMHMM-2.0/] *(74)*. In addition, the subcellular localization and presence of signal peptides can be assessed using PSORTb [http://www.psort.org/psortb/] *(75)* and SignalP [http://www.cbs.dtu.dk/services/SignalP/] *(76)*, respectively.

Table 6.1
Predictive power of online gene recognition tools. The sequence of *Salmonella* phage ε34 was annotated using a combination of Kodon software (Applied Maths, Austin, TX, USA) and BLASTX analyses. A total of 71 ORFs were defined

Software	URL	Model	No. of correctly annotated	No. with incorrect 5′ end	Miss-annotated	No. of Missing
AMIGene	1	STLT2	43	7	7	21
EasyGene	3	ST	49	2	4	20
FGENESB	4	ECK12	47	3	10	21
GeneHacker	5	EC	20	6	4	45
GeneMark	6	STLT2	56	7	3	8
Generation*	7	STLT2	12	1	6	58
Glimmer 2.02	8	ECK12	17	5	11	49
Glimmer 2.13	9	STLT2	40	7	5	24

EC: *Escherichia coli*; ECK12: *E. coli* K12; ST: *Salmonella* Typhi; STLT2: *Salmonella* Typhimurium LT2.
*This URL also provides Glimmer analysis; identifying considerably more ORFs.
Key to URL addresses:
1. http://www.genoscope.cns.fr/agc/tools/amiga/Form/ form.php
2. http://www.cbs.dtu.dk/services/EasyGene/
3. http://www.softberry.com/berry.phtml?topic=fgenesb& group=programs&subgroup=gfindb
4. http://www-btls.jst.go.jp/GeneHacker/
5. http://opal.biology.gatech.edu/GeneMark/gmhmm2_prok. cgi
6. http://compbio.ornl.gov/GP3/pro.html
7. http://nbc11.biologie.uni-kl.de/fs.cgi?main=http:nbc11.biologie.uni-kl.de/annotation_suite_tutorial/http://pathogene.swmed.edu/

8 Genome Comparisons

A variety of excellent tools are available for online analysis or downloading which can be used to compare phage-sized genomes. These function at the level of nucleotide or protein comparisons, and include at the simplest level dotplot comparison tools and progress to the more complex: ACT [Artemis Comparison Tool; http://www.sanger.ac.uk/Software/ACT/] *(77)*, MAUVE [http://gel.ahabs.wisc.edu/mauve/index.php] *(78)*, CoreGenes [http://binf.gmu.edu:8080/CoreGenes2.0] *(79)*, and GeneOrder (http://binf.gmu.edu:8080/GeneOrder3.0/) *(80)*. In the following sections, the utility of each of these programs will be discussed.

8.1 Dotplots

Sequence similarity between two genomes discovered using the BlastZ algorithm *(81)* can be visually represented in two formats:

PIP (percent identity plot, **Fig. 6.10**) or by dot-matrix plots (dotplots, **Fig. 6.11**). In the latter case, the two sequences are represented on the axes and where a region of sequence similarity exists a dot is placed on the image. If sufficient similarity exists the dots coalesce into a line providing one with an easy visualization of regions of sequence similarity and dissimilarity.

These figures, comparing coliphages HK022 with HK97 are typical of the results with dotplots obtained with phage DNA showing regions of homology interspersed with regions showing no sequence similarity. They were generated using zPicture at http://zpicture.dcode.org/(Comparative Genomics Center at Lawrence Livermore Laboratory) which is a variant of PipMaker *(81, 82)*. rVISTA at http://genome.lbl.gov/vista/mvista/submit.html also provides excellent PIP plots *(83)*. We would also recommend Java Dot Plot Alignments (JDotter) at Viral Bioinformatics Resource Center (http://athena.bioc.uvic.ca/workbench.php?tool=jdotter& db=) since it permits one to vary the alignment stringency.

8.2 ACT

The Artemis Comparison Tool (ACT) is a pairwise comparison tool for comparing any DNA sequences up to, an including, whole genomes *(77)*. It is built with the same components used in Artemis, so as well as sharing many tools and sequence views, it has a similar look and feel. Also, a separate Artemis window can be launched from within ACT to display and edit any one of the sequences loaded.

ACT can compare two or more sequences. The sequences are displayed one above the other and the regions of similarity are linked by red and blue bands (*see* **Fig. 6.12**). Red bands indicate forward matches and the blue bands represent reverse matches. These bands display different intensities of coloring depending on the strength of the matches they link; the higher the color intensity, the stronger the match between regions. When a region is selected or highlighted it turns yellow and the percentage of similarity and score are displayed on the top left corner of the com-

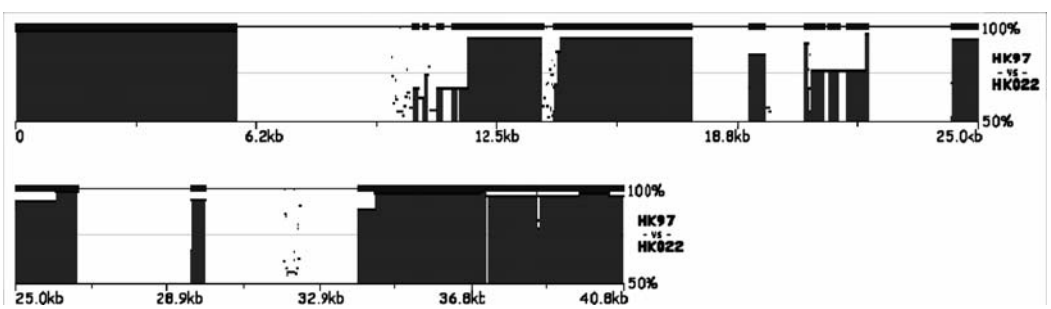


Fig. 6.10. Percent nucleotide identity plot comparing the genomes of coliphages HK97 and HK022. This reveals one of the defining aspects of phage genomes—their mosaic nature, that is, blocks of homology are separated by blocks which bear no sequence similarity.

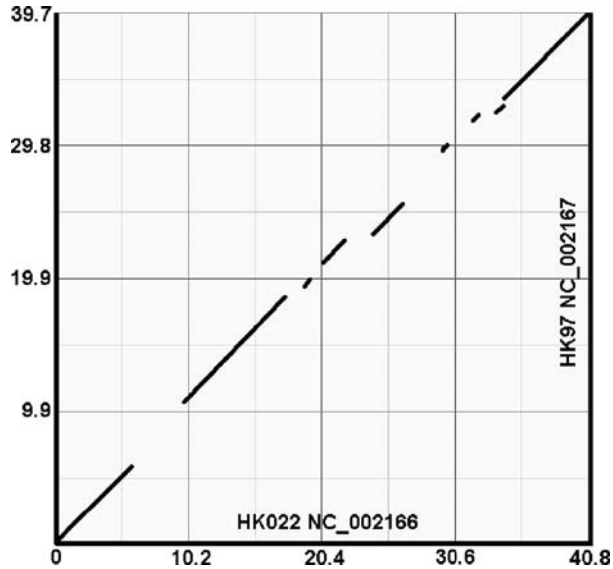


Fig. 6.11. Dotplot comparison of the nucleotide sequences of coliphages HK97 and HK022. The lines have been enhanced for this publication. These results indicate that the two genomes are collinear.

parison file window. These regions that match can be calculated from BLASTN or TBLASTX, or any other comparison program that will generate output in the required format.

The comparison files can be generated either locally or using web tools such as WebACT (http://www.webact.org/WebACT/home) *(84)* and Double ACT (http://www.hpa-bioinfotools.org.uk/pise/double_act.html). The DNA sequences to be compared do not have to contain any annotation for the comparison files to be generated. ACT accepts the same sequence formats (EMBL, GenBank, GFF or Fasta) as Artemis.

The same features that have been described in the Artemis section can be found in ACT. So searching for regions of interest, CDSs or features can also be done separately for each of the sequences that have been loaded into ACT. G + C content, and other, graphs can be displayed separately for each sequence. Also regions of interest can be zoomed in and out to display base pair differences and up to whole genome regions. This highlights regions of similarity, insertions/deletions or rearrangements (**Figs. 6.13A** and **B**).

An additional tool in ACT is the ability to identify regions of differences. This can be done in two different ways. The first is by selecting features in non-matching regions (an option under the "Select" menu). Alternatively, features can be created spanning the non-matching regions.

Artemis and ACT are actively being developed. They have been found to be invaluable tools in the annotation and analy-

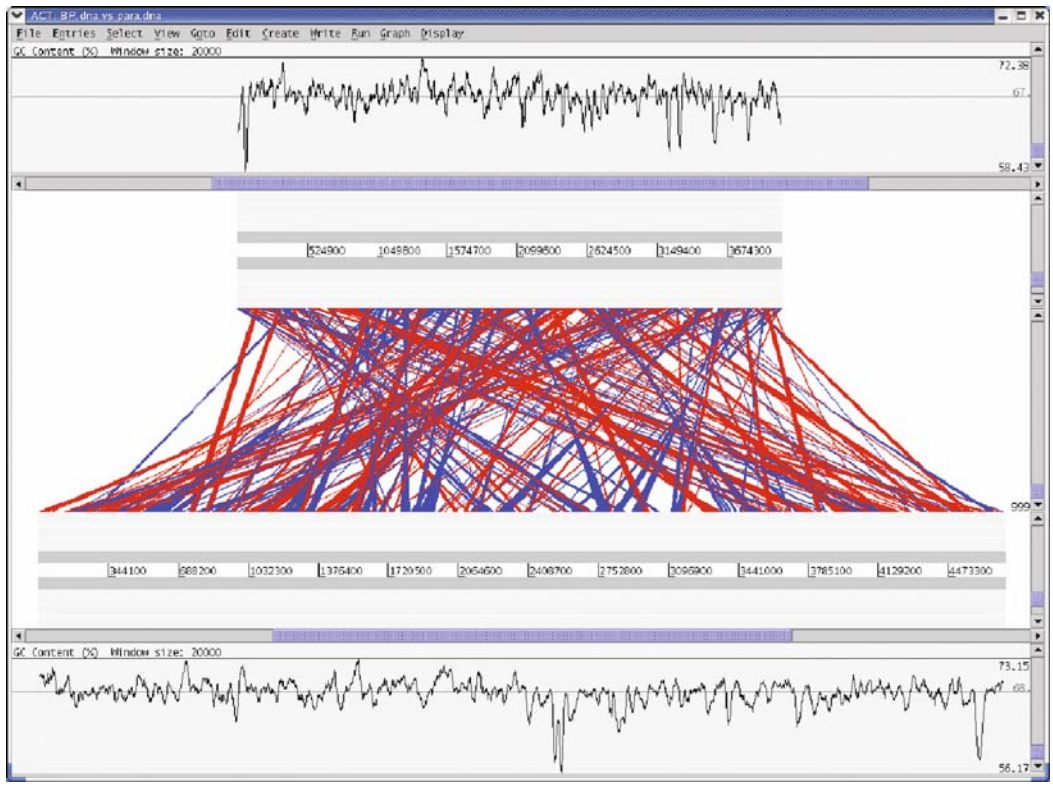


Fig. 6.12. Whole genome comparison between *Bordetella pertussis* (*top sequence*) and *Bordetella parapertussis* (*bottom sequence*). The comparison file shows the red and blue linking bands between the two genomes (see text for explanation). The GC contents of both genomes are also shown with a window size of 20 kb.

sis process. New and development releases are freely available for these software tools.

8.3 Mauve

Mauve is a freely-available, open-source tool for multiple genome sequence alignment and visualization *(78)*. It enables the rapid comparison of multiple genome sequences at the nucleotide level to identify orthologous regions, breakpoints of rearrangement, and lineage-specific sequence. In addition to creating a textual output of the multiple genome alignment in eXtended Multi-FastA format (XMFA), the Mauve alignment viewer provides interactive browsing of sequence similarity and annotated sequence features. Mauve works on Mac OS X, Windows, and Linux and is available in http://gel.ahabs.wisc.edu/mauve.

In order to compare phage genomes with Mauve, one must first acquire the subject sequences in either FastA or GenBank format. If GenBank files with annotated sequence features, such as CDS and RNA genes are used, the Mauve alignment viewer will display the annotated features. Each genome sequence must reside either in separate files or in

a single Multi-FastA or Multi-GenBank file. For the sake of exposition, we will align the genomes of *Staphylococcus* phage K, Twort, and G1 *(30, 85)*. These three genome sequences are available as a single Multi-GenBank file from http://gel.ahabs.wisc.edu/mauve/alignments/three_staph_phage.gbk.

To start the analysis, launch Mauve and select "Align..." from the "File" menu. A dialog box appears that can be used to select the input sequence files and set the alignment parameters (**Fig. 6.14**). To align the *Staphylococcus* phage, add the "three_staph_phage.gbk" file using the "Add sequence..." button. On Windows and Mac OS X, drag-and-drop can be used to place the "three_staph_phage.gbk" file into the "Align sequences..." dialog. Optionally, the location of the alignment output files can be specified by clicking the "..." button or typed in directly. Leaving the rest of the parameters unchanged

(A)

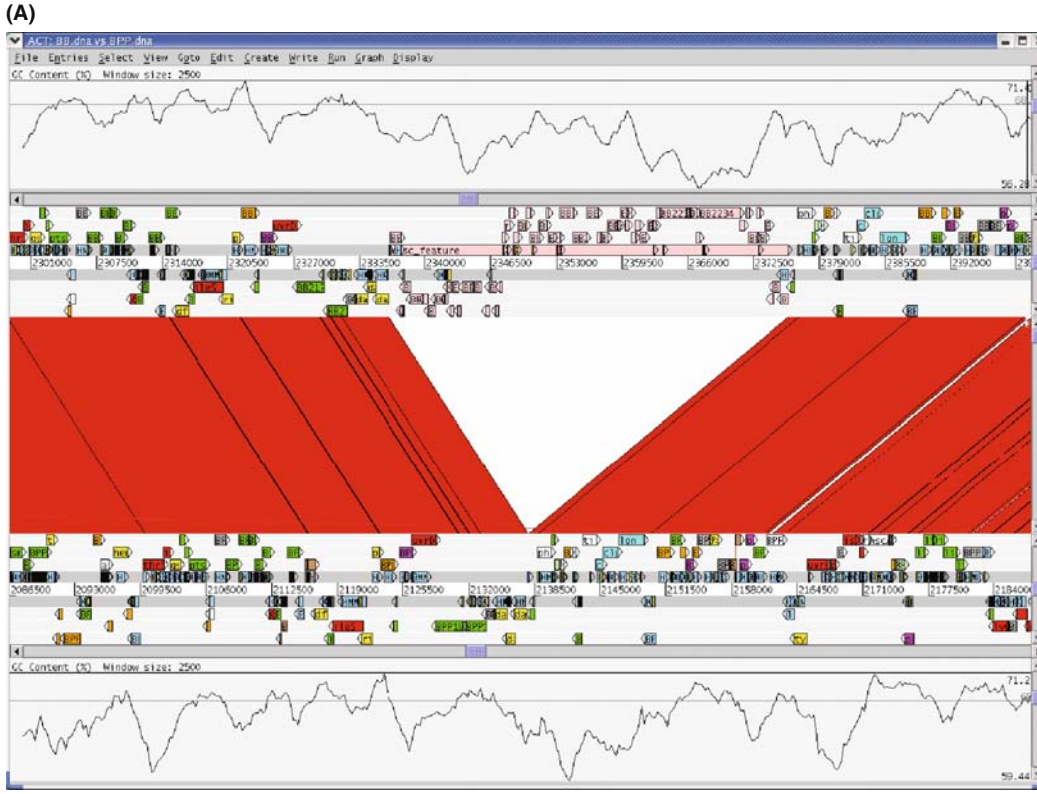


Fig. 6.13. **(A)** *Bordetella bronchi-septica (90)* prophage region displayed on the *top* compared with *B. parapertussis*, on the *bottom*, both with their respective GC content graphs (with a 2.5 kb window). Note the regions flanking the prophage share similarities in both genomes in which it seems to have been inserted. **(B)** A magnified view to show at the nucleotide level the insertion point for this prophage, which is a tRNA (*dark box*); a common chromosomal integration site for prophages.

(B)

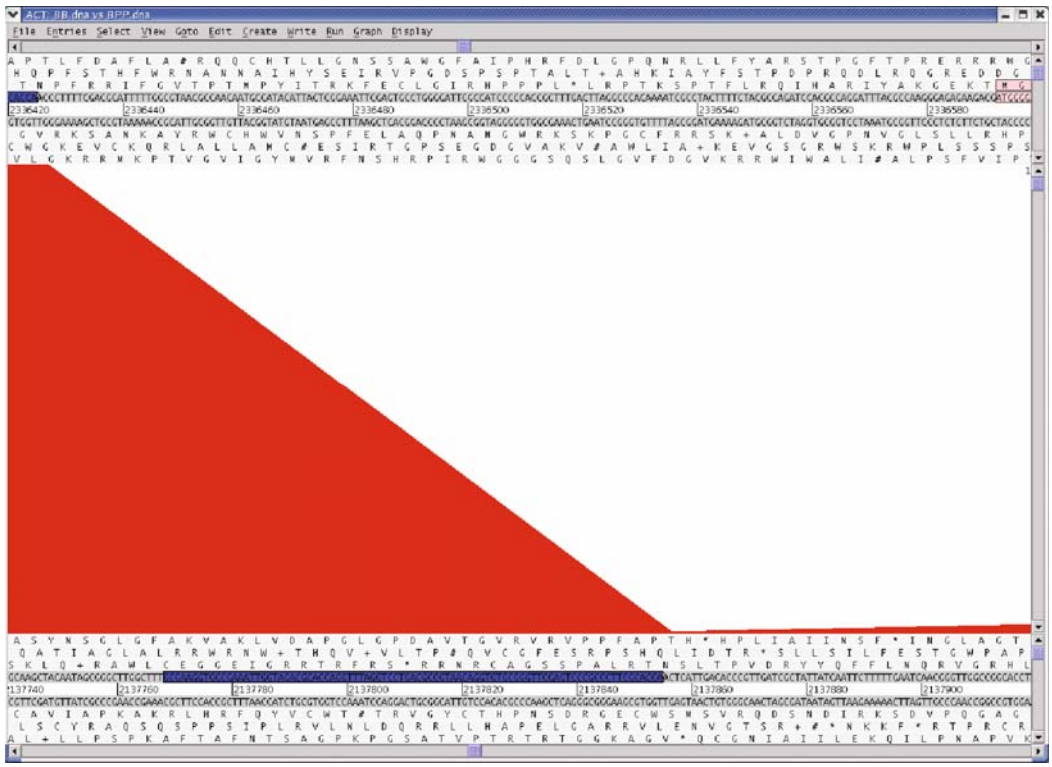


Fig. 6.13. (continued).

for the moment, we now perform the alignment by clicking the "Align. . ." button.

When the alignment computation completes, the alignment will load into the display window (**Fig. 6.15**). The alignment display shows each of the three genomes arranged horizontally, with a sequence similarity plot and annotated features for each genome. The display outlines putative orthologous regions with rounded rectangles that are linked among each genome. We refer to such homologs regions as locally collinear blocks (LCBs) because they are homologous blocks without any internal rearrangement. Within each LCB, the height of the colored plot indicates the amount of sequence similarity. When moving the mouse over one genome's similarity plot, a black line tracks the orthologous region in the other genomes, providing immediate feedback on how the regions are aligned. Clicking the similarity plot will vertically align the display on the orthologous region. Annotations appear immediately below each genome's similarity plot as rectangular boxes. The current releases of Mauve color code annotated CDS features as white, rRNA as red, misc_RNA as blue, and tRNA as green.

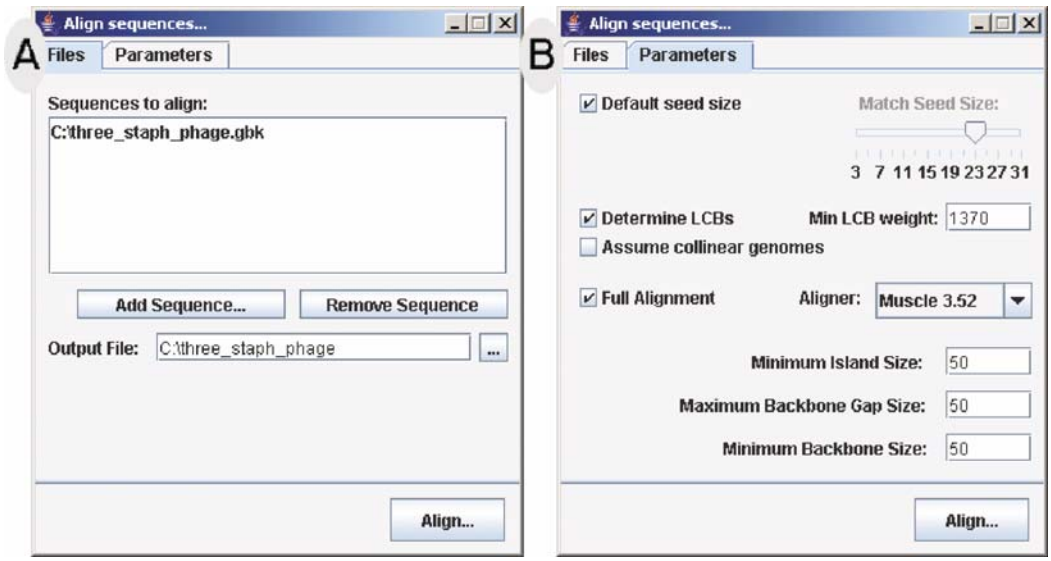


Fig. 6.14. The "Align sequences" interface in Mauve. This window permits the selection of files containing the genome sequences to be aligned and the configuration of various alignment parameters.

Fig. 6.15. The genomes of *Staphylococcus* phages K, Twort, and G1 as aligned by Mauve 1.2.3 with default parameter settings. Each of the three genomes is displayed horizontally with homologous locally collinear blocks (LCBs) outlined and connected with lines. Within each LCB, the height of the similarity plot indicates the average sequence similarity. Annotated genes are displayed as *rectangular white boxes* below each genome's similarity plot. The LCB weight slider (*top right*) can be used to adjust alignment parameters appropriately.

As described in Darling et al. *(78)*, Mauve uses a parameter called the minimum locally collinear block (LCB) weight to filter out regions of spurious homology. The default LCB weight parameter is typically too sensitive and causes spurious genome rearrangements to appear in the display. Thus, it is usually necessary to estimate a better LCB weight parameter. One way to estimate a "good" LCB weight is by first performing an alignment with default parameters, then using the LCB weight slider (top right in **Fig. 6.15**) to gradually increase the LCB weight until

the remaining regions appear to correspond to legitimate orthologous regions. For the present study of *Staphylococcus* phage, a good LCB weight parameter seems to be 1370 as it eliminates all small rearrangements leaving only the largest regions of homology. Once a good LCB weight has been determined, we can recomputed the alignment using this LCB weight. We once again select "Align. . ." from the "File" menu and this time enter 1370 for the LCB weight parameter (**Fig. 6.14B**).

The final alignment for *Staphylococcus* phages K, Twort, and G1 using a minimum LCB weight parameter of 1370 results in two locally collinear blocks. The corresponding XMFA alignment file can be downloaded from http://gel.ahabs.wisc.edu/mauve/alignments/three_staph_phage. alignment and accessed within Mauve using the "Open. . ." option of the "File" menu. Many comparative observations can be made about these phage genomes using the Mauve viewer. For example, by zooming in on the region between 93,500 and 101,500 of phage K and clicking to align the other genomes we can see a region annotated as a putative multi-part DNA polymerase gene. Holding the mouse over the gene box in phage K displays its annotated function, *see* **Fig. 6.16**. Interspersed within the putative DNA polymerase gene lie two other putative genes in the K and G1 genomes. Interestingly, the phage Twort genome appears to be missing the additional ORFs. By right-clicking on the gene annotations, a menu pops up that allows us to view the gene entry in NCBI Entrez to search for literature and additional information on these genes and their homologs in other species.

Like any automated alignment system, Mauve may align some sequence incorrectly. Fortunately, Mauve includes a graphical interface based on CINEMA-MX *(86)* to edit the sequence align-

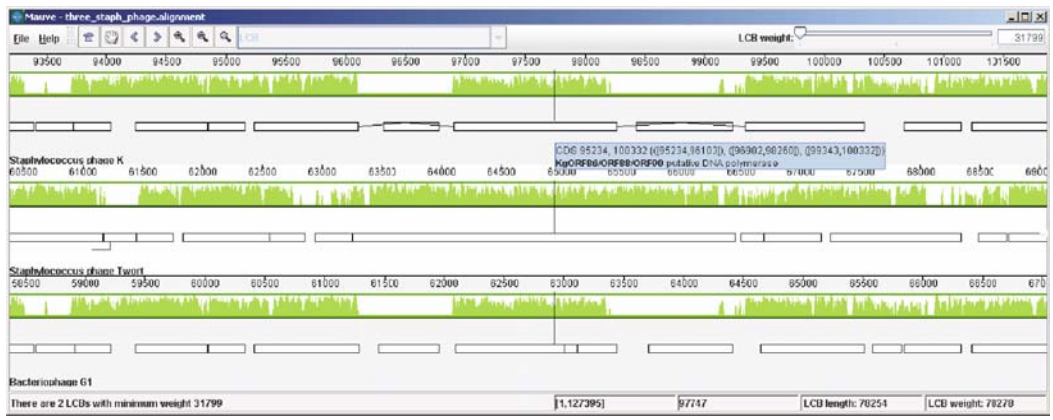


Fig. 6.16. A zoomed in view of a putative DNA polymerase gene in phage K. The gene is annotated as multi-part in phages K and G1, but not in phage Twort. Twort is missing two ORFs relative to K and G1 in this region. The annotation information appears when the mouse hovers over the gene. Zooming in the alignment can be accomplished using either the magnifying glass tools at top left or the Ctrl-Up Arrow and Ctrl-Down Arrow keyboard shortcuts.

ment and repair misaligned regions. The alignment editor can be activated by right-clicking on an LCB in the alignment viewer and selecting the "Edit this LCB..." item that appears in the pop-up menu. The Cinema-MX alignment editor appears momentarily with the alignment of the selected LCB. The alignment can be edited by dragging nucleotides to the left or right in each sequence. See Lord et al. *(86)* for more details. When editing has been completed, select "Save" from the "File..." menu and any changes will be saved to the XMFA alignment file. Mauve will automatically reload its display with the new alignment.

8.4 CoreGenes and GeneOrder

GeneOrder and CoreGenes are two related "on-the-fly" web-accessible software tools designed for the analysis and comparisons of genomes. They are located on the Department of Bioinformatics and Computational Biology server at George Mason University (http://www.binf.gmu.edu/genometools.html). These tools are coded in Java and are platform-independent. Both are based on a BLASTP tool (http://BLAST.wustl.edu) that allows a comparison of two (GeneOrder) or up to five (CoreGenes) genomes per session.

GeneOrder is designed to compare two whole genomes of up to approximately 2 megabases (Mb) each (GeneOrder3.0; http://binf.gmu.edu:8080/GeneOrder3.0). The outputs are a table and a graph, each comparing genes and genomic rearrangements; these in turn, highlight gene order and synteny between these two genomes, highlighting orthologous and paralogous genes, depending on the selected criteria. In brief, genome sequence entries are downloaded from GenBank via their accession numbers (see caveat below). The algorithm refers to one genome as the "reference" genome and to the other as the "query" genome. These files are parsed for annotated genes, which are then organized into a BLASTP format. The annotated "query genome" genes are then BLAST analyzed against the reference genome systematically. If the alignment score is equal to or exceeds the either default or custom-entered BLASTP high score threshold values, then the genes are paired and their gene numbers extracted and scored by the algorithm. Following the completion of the analysis, a table of these pairs, along with their high scores, is generated for GeneOrder *(87)*. These pairs link back to the individual gene entries in GenBank. An applet dotplot graph is generated in the current versions of GeneOrder. As an example, we have aligned the proteomes of coliphage T7 (NC_001604) with that of its relative *Pseudomonas* phage gh-1 (NC_004665) in **Fig. 6.17**.

CoreGenes is designed to compare two to five genomes, of sizes up to approximately 350,000 bases (CoreGenes1.0; http://binf.gmu.edu:8080/CoreGenes1.0), in a progressive analysis that yields a table of orthologous and paralogous genes, depending

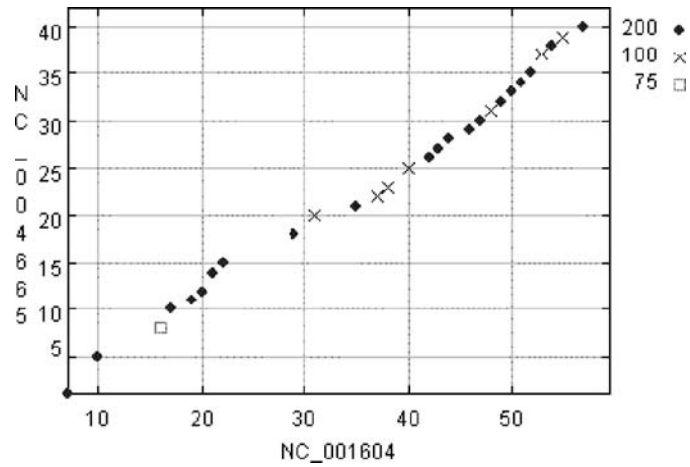


Fig. 6.17. GeneOrder plot of coliphage T7 (NC_001604) versus Pseudomonas phage gh-1 (NC_004665).

on the criteria selected. CoreGenes gives a table of related and potentially related genes across these genomes. CoreGenes uses the same algorithm for a progressive alignment of additional genomes. For this tool, the first two genomes are compared to generate a "consensus" genome that is then queried with a third genome. A subsequent "consensus" is generated and used as a reference genome for the next query genome, and so forth. A consequence is that similar genes that are not included in one of the early consensus genomes will not be included in subsequent analysis. This is one feature that will be addressed in upcoming releases. The above examples of T7-like phages may be run for illustration in groups of two to five genomes *(79)* (**Fig. 6.18**).

Since both software tools are based on a BLAST algorithm, the BLASTP high scores are used to parse the data. For GeneOrder, the user is offered matches in three ranges, to be entered in the threshold boxes: "highest," "high," and "low." The default values are set at [200–infinity) for highest, [100–200) for high, and [75–100) for low; these give a relative indication of identity and similarity, with highest representing identity matches, at one extreme, and low representing potentially similar matches that require further analysis, at the other extreme.

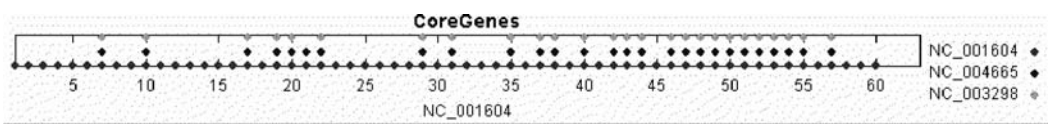


Fig. 6.18. CoreGene analysis of the proteomes of coliphage T7 (NC_001604; *dots on the bottom line*) with those of gh-1 (NC_004665); dots on the middle line and T3 (NC_003298, *dots on top line*).

For CoreGenes, one value acting as the BLASTP threshold high score may be entered. The default is 75, which will give many potential matches; again, all requiring further analysis to confirm. Higher stringency can be attained by changing this value to 100 or 200.

A caveat for both GeneOrder and CoreGenes is that the National Library of Medicine, which administers GenBank, apparently changes accession numbers without warning. Therefore, one must be sure the accession number is current and active. For example, NC_001406 used to be an accession number for human adenovirus serotype 5, but recently has changed to identifying a null genome. Also, rather than be despondent, one must recheck the genome entry in GenBank should an error be returned by either software. In some cases, the "Refresh" button of the browser needs to be used, after an error message is displayed.

Acknowledgements

Alexandre Lomsadze and Mark Borodovsky thank John Besemer for valuable comments. A.K. acknowledges the financial support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Grigoriev, A. 1999. Strand-specific compositional asymmetries in double-stranded DNA viruses. Virus Research *60*:1–19.
2. Casjens, S., D.A. Winn-Stapley, E.B. Gilcrease, R. Morona, C. Kuhlewein, J.E. Chua, P.A. Manning, and A.J. Clark. 2004. The Chromosome of *Shigella flexneri* Bacteriophage Sf6: Complete Nucleotide Sequence, Genetic Mosaicism, and DNA Packaging. Journal of Molecular Biology *339*:379–394.
3. Hu, F., K. Zhang, Y. Tan, X. Jin, J. Zhu, J. Huang, X. Rao, X. Shen, and X. Hu. 2003. Complete genome sequence of Pseudomonas aeruginosa bacteriophage PaP3. GenBank Accession Number NC_004466.
4. Mann, N.H., A. Cook, M. Clockie, and A. Millard. 2005. Sequence analysis of the genome of bacteriophage S-PM2. Cyanophage S-PM2. GenBank Accession Number NC_006820.
5. Lindell, D., M.B. Sullivan, Z.I. Johnson, A.C. Tolonen, F. Rohwer, and S.W. Crisholm. 2005. *Prochlorococcus* cyanophage genomes. Cyanophage P-SSM2. GenBank Accession Number NC_006883.
6. Nolan, J.M., V. Petrov, C. Bertrand, H.M. Krisch, and J.D. Karam. 2005. Comparative analysis of the *Aeromonas* bacteriophage 31 genome. GenBank Accession Number NC_007022.
7. Sibbald, M.J. and A.M. Kropinski. 1999. Transfer RNA genes and their significance to codon usage in the *Pseudomonas aeruginosa* lamboid bacteriophage D3. Canadian Journal of Microbiology *45*:791–796.
8. Miller, E.S., J.F. Heidelberg, J.A. Eisen, W.C. Nelson, A.S. Durkin, A. Ciecko, T.V. Feldblyum, O. White, et al. 2003. Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage. Journal of Bacteriology *185*:5220–5233.
9. Miller, E.C., E. Kutter, G. Mosig, F. Arisaka, T. Kunisawa, and W. Rüger. 2003. Bacteriophage T4 genome. Microbiology and Molecular Biology Reviews *67*: 86–156.
10. Dodd, I.B. and J.B. Egan. 2005. Bacteriophage 186 complete genome. Enterobacte-

ria phage 186. GenBank Accession Number NC_001317.
11. Lobocka, M.B., D.J. Rose, G. Plunkett, III, M. Rusin, A. Samojedny, H. Lehnherr, M.B. Yarmolinsky, and F.R. Blattner. 2004. Genome of bacteriophage P1. Journal of Bacteriology *186*:7032–7068.
12. Ksenzenko, V.N., A.V. Kaliman, A.I. Krutilina, and M.G. Shlyapnikov. 200. Bacteriophage T5 complete genome. Enterobacterial phage T5. GenBank Accession Number NC_005859.
13. Sriranngana than, N., J.M. Whichard, F.W. Pierson, and V. Kapur. 2005. Bacteriophage Felix O1: Genetic characterization. GenBank Accesion Number NC_005282.
14. Smith, M.C., R.N. Burns, S.E. Wilson, and M.A. Gregory. 1999. The complete genome sequence of the *Streptomyces* temperate phage straight ϕC31: evolutionary relationships to other viruses. Nucleic Acids Research *27*:2145–2155.
15. Gregory, M.A., R. Till, and M.C. Smith. 2003. Integration site for *Streptomyces* phage ϕBT1 and development of site-specific integrating vectors. Journal of Bacteriology *185*: 5320–5323.
16. Pedulla, M.L., M.E. Ford, J.M. Houtz, T. Karthikeyan, C. Wadsworth, J.A. Lewis, D.Jacobs-Sera, J. Falbo, et al. 2003. Origins of highly mosaic mycobacteriophage genomes. Cell *113*: 171–182.
17. Ford, M.E., G.J. Sardis, A.E. Belanger, R.W. Hendrix, and G.F. Hatfull. 1998. Genome structure of mycobacteriophage D29: Implications for phage evolution. Journal of Molecular Biology *279*:143–164.
18. Zimmer, M., E. Sattelberger, R.B. Inman, R. Calendar, and M.J. Loessner. 2003. Genome and proteome of *Listeria monocytogenes* phage PSA: an unusual case for programmed +1 translational frameshifting in structural protein synthesis. Molecular Microbiology *50*:303–317.
19. Lowe, T.M. and S.R. Eddy. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research *25*:955–964.
20. Laslett, D., Canback, B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Research *32(1)*: 11–16.
21. Newton, G.J., C. Daniels, L.L. Burrows, A.M. Kropinski, A.J. Clarke, and J.S. Lam. 2001. Three-component-mediated serotype conversion in *Pseudomonas aeruginosa* by bacteriophage D3. Molecular Microbiology *39*:1237–1247.
22. Lindell, D., M.B. Sullivan, Z. I. Johnson, A.C. Tolonen, F. Rohwer, and S. W. Chisholm. 2004. Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. Proceedings of the National Academy of Sciences of the United States of America *101*:11013–11018.
23. Villegas, A. and A.M. Kropinski. 2008. An analysis of initiation codon utilization in the domain Bacteria – concerms about the quality of bacterial genome annotation Microbiology *154*: 2559–2561.
24. Shine, J. and L. Dalgarno. 1975. Terminal-sequence analysis of bacterial ribosomal RNA. Correlation between the 3′-terminal-polypyrimidine sequence of 16-S RNA and translational specificity of the ribosome. European Journal of Biochemistry *57*:221–230.
25. Shine, J. and L. Dalgarno. 1974. The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. Proceedings of the National Academy of Sciences of the United States of America *71*:1342–1346.
26. Farinha, M.A. and A.M. Kropinski. 1997. Overexpression, purification, and analysis of the *c1* repressor protein of *Pseudomonas aeruginosa* bacteriophage D3. Canadian Journal of Microbiology *43*:220–226.
27. Hendrix, R.W. 2002. Bacteriophages: evolution of the majority. Theoretical Population Biology *61*:471–480.
28. Hendrix, R.W., M.C. Smith, R.N. Burns, M.E. Ford, and G.F. Hatfull. 1999. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proceedings of the National Academy of Sciences of the United States of America *96*:2192–2197.
29. Chibani-Chennoufi, S., C. Canchaya, A. Bruttin, and H. Brussow. 2004. Comparative genomics of the T4-Like *Escherichia coli* phage JS98: implications for the evolution of T4 phages. Journal of Bacteriology *186*: 8276–8286.
30. O'Flaherty, S., A. Coffey, R.Edwards, W. Meaney, G.F. Fitzgerald, and R.P. Ross. 2004. Genome of staphylococcal phage K: a new lineage of Myoviridae infecting gram-positive bacteria with a low G + C content. Journal of Bacteriology *186*: 2862–2871.
31. Scholl, D. and C. Merril. 2005. The Genome of bacteriophage K1F, a T7-Like phage that has acquired the ability to replicate on K1 strains of *Escherichia coli*. Journal of Bacteriology *187*:8499–8503.
32. Bonocora, R.P. and D.A. Shub. 2004. A self-splicing group I intron in DNA polymerase genes of T7-like bacteriophages. Journal of Bacteriology *186*:8153–8155.

33. Foley, S., A. Bruttin, and H. Brussow. 2000. Widespread distribution of a group I intron and its three deletion derivatives in the lysin gene of *Streptococcus thermophilus* bacteriophages. Journal of Virology *74*: 611–618.
34. Nelson, D., R. Schuch, S. Zhu, D.M. T scherne, and V.A. Fischetti. 2003. Genomic sequence of C1, the first streptococcal phage. Journal of Bacteriology *185*:3325–3332.
35. Seegers, J.F., G.S. Mc, M.O'Connell-Motherway, E.K. Arendt, G.M. van de, M. Creaven, G.F. Fitzgerald, and S.D. van. 2004. Molecular and transcriptional analysis of the temperate lactococcal bacteriophage Tuc2009. Virology *329*:40–52.
36. van, S.D., H. Karsens, J. Kok, P. Terpstra, M.H. Ruiters, G. Venema, and A. Nauta. 1996. Sequence analysis and molecular characterization of the temperate lactococcal bacteriophage r1t. Molecular Microbiology *19*: 1343–1355.
37. Landthaler, M. and D.A. Shub. 2003. The nicking homing endonuclease I-*Bas*I is encoded by a group I intron in the DNA polymerase gene of the *Bacillus thuringiensis* phage Bastille. Nucleic Acids Research *31*:3071–3077.
38. Lazarevic, V., B. Soldo, A. Dusterhoft, H. Hilbert, C. Mauel, and D. Karamata. 1998. Introns and intein coding sequence in the ribonucleotide reductase genes of *Bacillus subtilis* temperate bacteriophage SPbeta. Proceedings of the National Academy of Sciences of the United States of America *95*: 1692–1697.
39. Mann, N.H., M.R. Clokie, A. Millard, A. Cook, W.H. Wilson, P. J. Wheatley, A. Letarov, and H. M. Krisch. 2005. The genome of S-PM2, a "photosynthetic" T4-type bacteriophage that infects marine *Synechococcus* strains. Journal of Bacteriology *187*: 3188–3200.
40. Landthaler, M. and D.A. Shub. 1999. Unexpected abundance of self-splicing introns in the genome of bacteriophage Twort: introns in multiple genes, a single gene with three introns, and exon skipping by group I ribozymes. Proceedings of the National Academy of Sciences of the United States of America *96*:7005–7010.
41. Condron, B.G., J.F. Atkins, and R.F. Gesteland. 1991. Framshifting in gene 10 of bacteriophage T7. Journal of Bacteriology *173*:6998–7003.
42. Levin, M.E., R.W. Hendrix, and S.R. Casjens. 1993. A programmed translational frameshift is required for the synthesis of a bacteriophage lambda tail assembly protein. Journal of Molecular Biology *234*: 124–139.
43. Christie, G.E., L.M. Temple, B.A. Bartlett, and T.S. Goodwin. 2002. Programmed translational frameshift in the bacteriophage P2 FETUD tail gene operon. Journal of Bacteriology *184*:6522–6531.
44. Kolla, V., M. Chakravorty, B. Pandey, S.M. Srinivasula, A. Mukherjee, and G. Litwack. 2000. Synthesis of a bacteriophage MB78 late protein by novel ribosomal frameshifting. Gene *254*:209–217.
45. Farabaugh, P.J. 1996. Programmed translational frameshifting. Annual Review of Genetics *30*:507–528.
46. Reeder, J. and R. Giegerich. 2004. Design implementation and evaluation of a practical pseudoknots folding algorithm based upon thermodynamics. BMC Bioinformatics *5*: 104–115.
47. Kropinski, A.M., M. Hayward, M.D. Agnew, and K.F. Jarrell. 2005. The genome of BCJA1c: a bacteriophage active against the alkaliphilic bacterium, *Bacillus clarkii*. Extremophiles *9*:99–109.
48. Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. Journal of Molecular Biology *215*:403–410.
49. Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. A. Rajandream, and B.Barrell. 2000. Artemis: sequence visualization and annotation. Bioinformatics *16*:944–945.
50. Besemer, J. and M. Borodovsky. 2005. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Research *33*:W451–W454.
51. Borodovsky, M. and J. McIninch. 1993. GeneMark: parallel gene recognition for both DNA strands. Computers & Chemistry *17*:123–133.
52. Lukashin, A. and M. Borodovsky. 1998. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Research *26*:1107–1115.
53. Besemer, J., A. Lomsadze, M. Borodovsky, J. Besemer, A. Lomsadze, and M. Borodovsky. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Research *29*:2607–2618.
54. Mills, R., M.Rozanov, A. Lomsadze, T. Tatusova, M. Borodovsky, R. Mills, M. Rozanov, A. Lomsadze, et al. 2003. Improving gene annotation of complete viral genomes. Nucleic Acids Research *31*: 7041–7055.

55. Besemer, J. and M. Borodovsky. 1999. Heuristic approach to deriving models for gene finding. Nucleic Acids Research *27*:3911–3920.
56. Kattenhorn, L.M., R. Mills, M. Wagner, A. Lomsadze, V. Makeev, M. Borodovsky, H.L. Ploegh, B.M. Kessler, et al. 2004. Identification of proteins associated with murine cytomegalovirus virions. Journal of Virology *78*:11187–11197.
57. Van Domselaar, G.H., P. Stothard, S. Shrivastava, J.A. Cruz, A. Guo, X. Dong, P. Lu, D. Szafron, et al. 2005. BASys: a web server for automated bacterial genome annotation. Nucleic Acids Research *33*:W455–W459.
58. Delcher, A.L., D. Harmon, S. Kasif, O. White, and S.L. Salzberg. 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Research *27*:4636–4641.
59. Larsen, T.S. and A. Krogh. 2003. EasyGene – a prokaryotic gene finder that ranks ORFs by statistical significance. BMC Bioinformatics *4*:21.
60. Bairoch, A., R. Apweiler, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, et al. 2005. The Universal Protein Resource (UniProt). Nucleic Acids Research *33*:D154–D159.
61. Bateman, A., L. Coin, R. Durbin, R.D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, et al. 2004. The Pfam protein families database. Nucleic Acids Research *32 Database issue*:D138–D141.
62. Hulo, N., C.J. Sigrist, S. Le, V, P.S. Langendijk-Genevaux, L. Bordoli, A. Gattiker, C.E. De, P. Bucher, and A. Bairoch. 2004. Recent improvements to the PROSITE database. Nucleic Acids Research *32*:D134–D137.
63. McGuffin, L.J., K. Bryson, and D.T. Jones. 2000. The PSIPRED protein structure prediction server. Bioinformatics *16*:404–405.
64. Bocs, S., S. Cruveiller, D. Vallenet, G. Nuel, C. Medigue, S. Bocs, S. Cruveiller, D. Vallenet, et al. 2003. AMIGene: Annotation of MIcrobial Genes. Nucleic Acids Research *31*: 3723–3726.
65. Yada, T. and M. Hirosawa. 1996. Detection of short protein coding regions within the cyanobacterium genome: application of the hidden Markov model. DNA Research *3*:355–361.
66. Salzberg, S.L., A.L. Delcher, S. Kasif, and O. White. 1998. Microbial gene identification using interpolated Markov models. Nucleic Acids Research *26*:544–548.
67. Ng, K.W., J. Lawson, H.R. Garner, K.w. Ng, J. Lawson, and H.R. Garner. 2004. PathoGene: a pathogen coding sequence discovery and analysis resource. BioTechniques *37*:218–2.
68. Altschul, S.F. and E.V. Koonin. 1998. Iterated profile searches with PSI-BLAST – a tool for discovery in protein databases. Trends in Biochemical Sciences *23*:444–447.
69. Marchler-Bauer, A., J.B. Anderson, C. DeWeese-Scott, N.D. Fedorova, L.Y. Geer, S. He, D.I. Hurwitz, J.D. Jackson, et al. 2003. CDD: a curated Entrez database of conserved domain alignments. Nucleic Acids Research *31*:383–387.
70. Letunic, I., R.R. Copley, S. Schmidt, F.D. Ciccarelli, T. Doerks, J. Schultz, C.P. Ponting, and P. Bork. 2004. SMART 4.0: towards genomic data integration. Nucleic Acids Research *32 Database issue*:D142–D144.
71. Tatusov, R.L., N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, D.M. Krylov, R. Mazumder, et al. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics *4*:41.
72. Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez. 2005. InterProScan: protein domains identifier. Nucleic Acids Research *33*:W116–W120.
73. Mulder, N.J., R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bradley, P. Bork, et al. 2005. InterPro, progress and status in 2005. Nucleic Acids Research *33*:D201–D205.
74. Sonnhammer, E.L.L., G. von Heijne, and A. Krogh. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences., p. 175–182. *In* J. Glasgow, Littlejohn T., F. Major, R. Lathrop, D. Sankoff, and C. Sensen (Eds.), Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, CA.
75. Rey, S., M.Acab, J.L. Gardy, M.R. Laird, K. deFays, C. Lambert, and F.S. Brinkman. 2005. PSORTdb: a protein subcellular localization database for bacteria. Nucleic Acids Research *33*:D164–D168.
76. Bendtsen, J.D., H. Nielsen, H.G. von, and S. Brunak. 2004. Improved prediction of signal peptides: SignalP 3.0. Journal of Molecular Biology *340*:783–795.
77. Carver, T.J., K.M. Rutherford, M. Berriman, M.-A. Rajandream, B.G. Barrell, and J. Parkhill. 2005. ACT: the Artemis comparison tool. Bioinformatics *21*:3422–3423.
78. Darling, A.C., B. Mau, F.R. Blattner, and N.T. Perna. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Research *14*:1394–1403.
79. Zafar, N., R.Mazumder, and D. Seto. 2002. CoreGenes: a computational tool for identifying and cataloging "core" genes in a set of small genomes. BMC Bioinformatics *3*:12.

80. Celamkoti, S., S. Kundeti, A. Purkayastha, R. Mazumder, C. Buck, and D. Seto. 2004. GeneOrder3.0: software for comparing the order of genes in pairs of small bacterial genomes. BMC Bioinformatics *5*:52.
81. Schwartz, S., Z. Zhang, K.A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, and W. Miller. 2000. PipMaker – a web server for aligning two genomic DNA sequences. Genome Research *10*:577–586.
82. Ovcharenko, I., G.G. Loots, R.C. Hardison, W. Miller, and L. Stubbs. 2004. zPicture: dynamic alignment and visualization tool for analyzing conservation profiles. Genome Research *14*:472–477.
83. Loots, G.G. and I. Ovcharenko. 2004. rVISTA 2.0: evolutionary analysis of transcription factor binding sites. Nucleic Acids Research *32*:W217–W221.
84. Abbott, J.C., D.M. Aanensen, K. Rutherford, S. Butcher, and B. G. Spratt. 2005. WebACT – an online companion for the Artemis Comparison Tool. Bioinformatics *21*: 3665–3666.
85. Kwan, T., J. Liu, M. DuBow, P. Gros, and J. Pelletier. 2005. The complete genomes and proteomes of 27 *Staphylococcus aureus* bacteriophages. Proceedings of the National Academy of Sciences of the United States of America *102*:5174–5179.
86. Lord, P.W., J.N. Selley, and T.K. Attwood. 2002. CINEMA-MX: a modular multiple alignment editor. Bioinformatics *18*:1402–1403.
87. Mazumder, R., A. Kolaskar, and D. Seto. 2001. GeneOrder: comparing the order of genes in small genomes. Bioinformatics *17*:162–166.
88. Cerdeno-Tarraga, A.M., S. Patrick, L.C. Crossman, G. Blakely, V. Abratt, N. Lennard, I. Poxton, B. Duerden, et al. 2005. Extensive DNA inversions in the *B. fragilis* genome control variable gene expression. Science *307*:1463–1465.
89. Cerdeno-Tarraga, A.M., A. Efstratiou, L.G. Dover, M.T. Holden, M. Pallen, S.D. Bentley, G.S. Besra, C. Churcher, et al. 2003. The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129. Nucleic Acids Research *31*:6516–6523.
90. Parkhill, J., M. Sebaihia, A. Preston, L.D. Murphy, N. Thomson, D.E. Harris, M.T. Holden, C.M. Churcher, et al. 2003. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. Nature Genetics *35*:32–40.

Chapter 7

Determining DNA Packaging Strategy by Analysis of the Termini of the Chromosomes in Tailed-Bacteriophage Virions

Sherwood R. Casjens and Eddie B. Gilcrease

Abstract

Tailed-bacteriophage virions contain a single linear dsDNA chromosome which can range in size from about 18 to 500 kbp across the known tailed-phage types. These linear chromosomes can have one of several known types of termini as follows: cohesive ends (5′- or 3′-single-strand extensions), circularly permuted direct terminal repeats, short or long exact direct terminal repeats, terminal host DNA sequences, or covalently bound terminal proteins. These different types of ends reflect differing DNA replication strategies and especially differing terminase actions during DNA packaging. In general, complete genome sequence determination does not by itself elucidate the nature of these ends, so directed experimental analysis is usually required to understand the nature of the virion chromosome ends. This chapter discusses these methods.

Key words: Terminase, tailed-phages, cohesive ends, terminal redundancy, DNA packaging.

1 Introduction

1.1 General Background

The tailed-bacteriophage virions all contain single, linear dsDNA molecules that are packaged into a procapsid by similar DNA translocase molecular motors; however, their DNA replication strategies and the resulting nature of ends of the packaged DNAs are not all the same. Their virion DNAs have six well-studied types of termini which are characterized by the presence of (i) single-stranded cohesive ends, (ii) circularly permuted direct terminal repeats, (iii) short, several hundred base pairs exact (non-permuted) direct terminal repeats, (iv) long, several thousand base pairs exact (non-permuted) direct terminal repeats, (v) terminal host DNA sequences, and (vi) covalently bound terminal proteins (**Table 7.1**). Five of the above six types of

Martha R. J. Clokie, Andrew M. Kropinski (eds.), *Bacteriophages: Methods and Protocols, Volume 2: Molecular and Applied Aspects, vol. 502*, © 2009 Humana Press, a part of Springer Science+Business Media
DOI 10.1007/978-1-60327-565-1_7 Springerprotocols.com

Table 7.1
Termini of tailed-phage virion DNAs.

Terminus type	Prototype phage	Replication strategy
Cohesive ends		
5′-single-strand extension	λ P2	Rolling circle → concatemer Circle→circle
3′-single-strand extension	HK97	Rolling circle →concatemer*
Circularly permuted direct terminal repeats†		
	T4	Complex→concatemer
	P22	Rolling circle→concatemer
	P1	Rolling circle→concatemer
Host DNA at termini		
	Mu	Duplicative transposition into host DNA
Exact direct terminal repeats		
Short (few hundred base pairs)	T7	Linear→concatemer
Long (thousands of base pairs)	SPO1	Complex→concatemer
	T5	Complex→concatemer
Covalent terminal protein		
	ϕ29	Protein-primed linear→linear

†Individual virions of these phages have chromosomes that terminate at many different places on the genome sequence, and the length of the terminal repeat varies among individual virions (see text).
*Genomic analysis predicts this mode of replication, but it has not been studied experimentally.

virion chromosome ends are generated by nucleolytic cleavage from replicating concatemeric or circular DNA molecules; only the phage genomes with terminal proteins do not require cleavage. These cleavages are in all cases tightly coupled to the DNA packaging process and are performed by a phage-encoded enzyme called "terminase," so named because this cleavage creates the termini of the virion DNA *(1, 2)*.

Most tailed-phage package DNA from concatemeric substrates that result from rolling-circle or more complex replication mechanisms. Among those studied, only the P2-like phages (subgroup of phage whose chromosomes have 5′-cohesive "COS" ends), the ϕ29-like phages whose chromosomes have covalently

bound terminal proteins, and the Mu-like phages whose DNA is integrated into host DNA, have packaging substrates that are monomeric; note that the first and last of these require that cleavage by terminase linearize the molecule for packaging and release the integrated phage DNA, respectively. The phages that package from concatemers typically engage in unidirectional "packaging series" on the DNA concatemers, where non-series-initiating packaging events begin at a DNA end generated by the previous event. This has been studied in most detail in mid-sized phages, such as P22 and λ, and the details of concatemer handling during packaging in the larger, more complex phages like T4 and SPO1 remain more poorly understood. **Figure 7.1** shows packaging series for four well-studied phages λ (COS ends), T7 (short direct terminal repeat or short DTR), P22 (terminally redundant and circularly permuted), and Mu (terminal host DNA).

The first event in such packaging series is recognition of the DNA by terminase, and then a double-strand cleavage is made at or near the packaging recognition site (typically called *pac* in headful packaging phages *(3)* and *cos* in cohesive end phages *(4)*). Only one of the two DNA ends created by this first (packaging series initiation) cleavage is threaded into a viral procapsid so that the packaging motor inserts DNA into the procapsid in *only one direction* from the cleavage point (**Fig. 7.1**). When DNA has filled the procapsid, a second nucleolytic cut (the "headful" cleavage) is made by terminase, which releases the packaged DNA from the concatemer, thus terminating the first packaging event. A second packaging event on that concatemer is then initiated by insertion of the unpackaged concatemer end created by the previous headful cleavage into a new procapsid. The second event is terminated like the first, with a headful cleavage (the second headful cleavage of the series). Subsequent packaging events then follow sequentially in the same manner as the second. Such *unidirectional packaging series* are usually two to five packaging events long, but can be upto 10 or more events long, depending on infection conditions *(5)*.

Below we discuss briefly how the different kinds of virion DNA ends result from different replication/terminase cleavage/packaging mechanisms. Demonstration of the presence of each of these types of ends requires specialized, directed analysis. Successful analysis of these ends requires more understanding of the various possible DNA end styles and how they are generated than it does technically difficult experimental analysis. Particular attention is paid here to phages whose genomes are completely sequenced, since at present this is very often the case for phages whose end structures are of interest. When phage genomes are sequenced by random shotgun sequencing (or even by primer walking on a phage chromosome template) *apparently* circular sequences are generated for circularly permuted and terminal

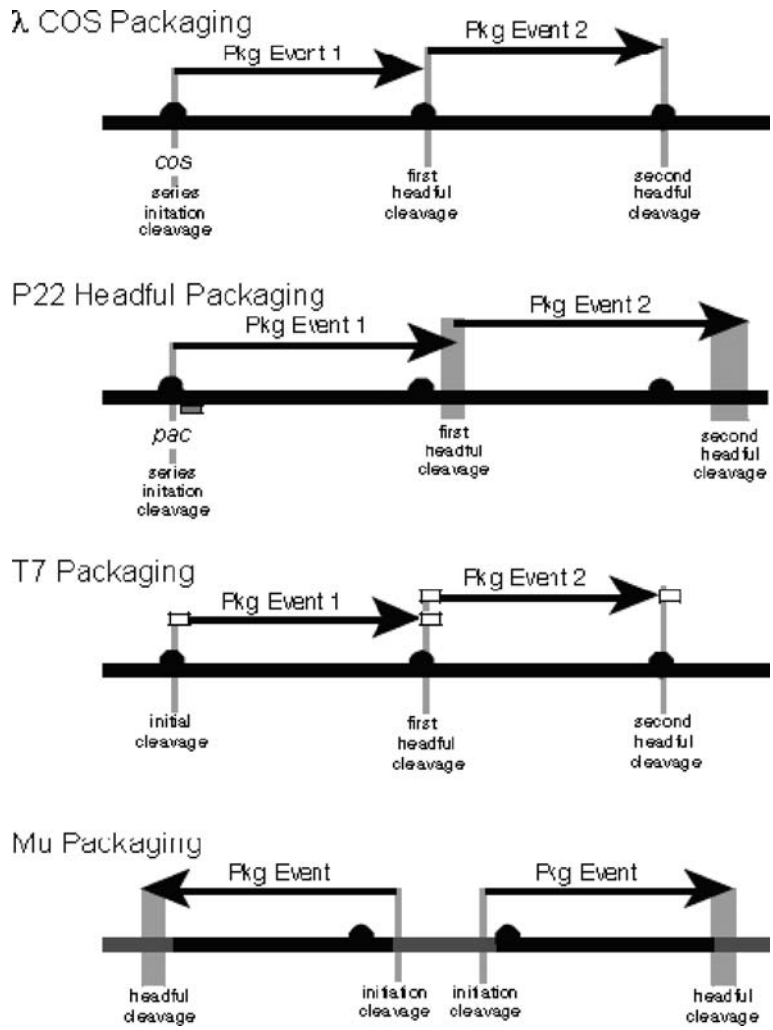


Fig. 7.1. Four tailed-phage DNA packaging strategies. Packaging strategies of phages λ, P22, T7, and Mu are shown diagrammatically. *Thick black horizontal lines* represent phage concatemeric DNAs, or, in the case of Mu, phage genomes that are integrated into the host bacteria's chromosome (the latter represented by *thick gray line*). *Black circles* mark the packaging recognition sites and *horizontal black arrows* represent individual packaging events. *Vertical black lines* indicate precise terminase cleavages and *vertical gray lines* indicate imprecise cleavages (see text). In each case except Mu, sequential series of packaging events occur, in which subsequent events (event 2 in figure) on the same concatemer molecule begin at the concatemer end created by the previous event (event 1); although only two successive events are shown, packaging series can in some cases be up to 10 or more events long. In phages λ and T7, each event begins and ends at a packaging recognition site, and in phage T7 the white rectangles show the region (the direct terminal repeat) that is duplicated in concert with packaging. In phage P22, the increasing width of the vertical gray boxes to the right, denotes the increased range of cleavage site locations as events proceed rightward. The small *gray horizontal rectangle* below the first P22 event is the optimal location of the Southern probe used to analyze pac fragments (see text).

exact direct repeat genomes, and even for cohesive end phages if the plasmid-cloned phage DNA inserts include cohesive ends that have been ligated together. Of course these are all *artificially circular* sequences, since *all* known tailed-phage virion chromosomes are linear. Our current knowledge of the mechanism of DNA packaging and injection suggests that covalently-circular virion DNA molecules will never be found in a tailed-phage, since the dsDNA must be threaded into the virion during packaging and out of the virion during injection through a narrow "portal" passage that will not accommodate two parallel dsDNAs simultaneously (which would be required if the chromosome were circular) *(6)*. Thus, even when the complete genome sequence has been determined, additional experiments are usually required to understand the true nature of the linear virion DNA.

1.2 Cohesive Ends

1.2.1 Best Studied Phages—λ, HK97, and P2

The two ends of cohesive end-containing phage chromosomes have protruding single-strands of identical length that are complementary to one another in sequence; upon injection these two ends anneal to each other, and each strand is closed by DNA ligase (of the host in those cases studied) to generate the covalently-closed circular molecule that serves as a template for DNA replication. Such cohesive ends can have either 5′- or 3′-protruding strands *(7,8,9)* and have been reported to be between 7 and 19 nucleotides in length in various phages (e.g., P2 has 19 nucleotide 5′-protruding strands *(10)* and HP1 has 7 nucleotide 5′-protruding strands *(11)*). Such ends are generated when the terminase makes *staggered, sequence-specific* cuts in the two DNA strands as it is being packaged. On a concatemer, a pair of staggered cuts (separated by the length of the eventual single-strand extension) generates the right end of one chromosome and the left end of the next chromosome to be packaged (except for the first and last cuts in a packaging series, where DNA one only one side of the cleavage is packaged; **Fig. 7.1**). Thus, the cohesive end termini of all individuals of a given COS phage are present at identical locations on the genome sequence.

DNAs with cohesive ends can be recognized by the ability of the opposite ends to anneal in the test tube, and the simplest way to detect such annealing is by the joining or "coherence" (hence the name *(7)*) of the two terminal restriction fragments. This annealing is most easily observed in agarose electrophoresis gels. Thus, if restricted DNA that contains cohesive ends is heated to a temperature that separates the cohesive ends but does not separate the strands of the remainder of the DNA (e.g., 75–80 °C) and cooled either slowly or rapidly. The cohesive-ended fragments will, under slow cooling conditions, anneal to one another and be visible as a larger band in such gels, but under fast cooling conditions, the two terminal bands do not have time to anneal with one another (**Fig. 7.2A**). We note that restriction enzyme-

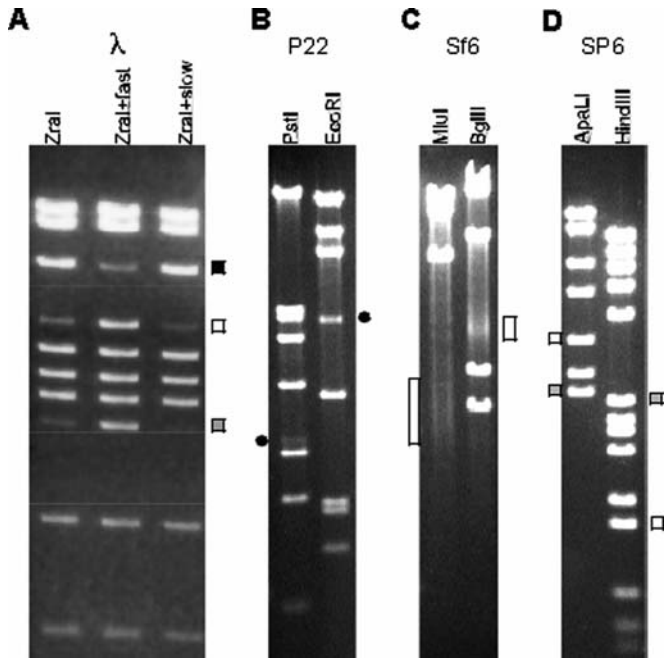


Fig. 7.2. Restriction enzyme generated fragments from tailed-phage virion DNA. DNA fragments were separated by 0.8% agarose gel electrophoresis and visualized by staining with ethidium bromide. The phage virion source of the DNA is indicated above each panel, and the restriction enzymes used are indicated above each lane. A. Phage λDNA after normal isolation and storage. DNA in the second and third lanes was heated to 75 °C for 15 min and then fast or slow cooled to room temperature, respectively, as described in the Methods. Terminal DNA fragments are indicated as follows: *white square*, left end fragment; *gray square*, right end fragment; *black square*, left and right end fragments annealed together by their cohesive ends. B. Phage P22 DNA. The pac fragments generated by PstI and EcoRI are indicated by *black circles* on the *left* and *right*, respectively. C. Phage Sf6 DNA. The locations of the diffuse pac fragment bands generated by imprecise packaging series initiation are indicated by *white rectangles* (*see* also **ref**. *(31)*). D. Phage SF6 DNA. Terminal DNA fragments are indicated as follows: *white square*, left end fragment; *gray square*, right end fragment.

generated ends typically have ≤ 5 bp single-strand protrusions, which are insufficient to keep two fragments together in such a gel. In some cases, for example *Bacillus subtilis* phage $\phi 105$, the seven nucleotide 3′-cohesive ends have been reported to associate unusually rapidly and addition of formamide and/or treatment with single-strand specific nucleases were required to separate the joined end fragments *(9)*.

This simple analysis can indicate whether cohesive ends are present, but does not determine the single-strand extension length or which strand protrudes. Typically, the exact nature of such ends is determined by running dideoxynucleotide sequencing reactions off of both ends of a native chromosome template DNA to determine exactly where the template ends at each terminus. The locations of the 5′-end of each strand can then be

deduced by comparison to sequences across closed, ligated ends *(12)*, and being aware that Taq DNA polymerase tends to add at least one non-templated A as it runs off the end of a template. The 3′-ends of each strand are more difficult to determine directly, and their positions are typically deduced by assuming, since the strand breaks at annealed cohesive ends are ligatable, that the two ends were generated by single-strand nicks as described above. The above determination is straightforward if one knows about where the ends are on the genome. The approximate locations of the terminase-generated ends (*cos* site) can be determined from the sizes of the termini-containing restriction fragments (in the above fast vs. slow cooling experiment) and a restriction map or the location of the restriction sites in the genome sequence. In addition, the location of the *cos* site is quite highly conserved, and in a large majority of the cases studied to date, it is within about one Kbp and transcriptionally upstream of the gene that encodes the small terminase subunit. Although the small subunit genes are very variable in sequence, the more highly conserved terminase large subunit and portal proteins and the (nearly always) highly stereotyped order of the phage head assembly genes can often allow an informed guess for the location of the small terminase subunit gene and hence the *cos* site *(13)*. We hasten to mention that exceptions to this generality do exist, which include the P2-like phages, *Mycobacterium tuberculosis* phages L5 and D29 *(14,15)* and *Lactococcus lactis* phages r1t and c2 *(16,17)*; in each of these cases, the head genes appear have an atypical order and in the latter four cases the small terminase gene is not yet recognizable. In such cases, the cohesive ends must be located by restriction mapping before they can be characterized in detail.

1.3 Headful Packaging

1.3.1 Best Studied Phages—P22, P1, SPP1, and T4

Phages that contain chromosomes that are terminally redundant and circularly permuted are called "headful packaging" phages *(18)*. Phage P22 is the best characterized headful packaging phage. In this case, a specific site is recognized on the replicated concatemer to initiate a packaging series *(19)*, but it is the available volume inside of the head, not DNA sequence, that determines the location of subsequent cleavages in the packaging series (**Fig. 7.1**) *(20,21,22)*. This packaging series initiation site is called *pac*, and this name is typically reserved for the site that terminase recognizes to begin a headful packaging series. The terminase makes a sequence-specific cleavage (see below) to initiate series, but has little sequence specificity at the following headful cleavages, that are made only when the procapsid (phage head precursor) is "full" of DNA *(23)*. The packaged DNA length in the headful packaging phages is typically between 102% and 110% the length of the genome sequence, so these chromosomes have direct terminal repeats that vary from 2% to 10% of the genome length in different phages. Upon infection, homologous

recombination between these direct, terminal repeats generates the circular genome that is the template for DNA replication. One consequence of this packaging strategy is that the ends of each successive packaging event in a series "move" along the genome sequence (rightward in **Fig. 7.1**) from those generated by the previous event by the length of the terminal repeat. The result of such packaging series is that the virion chromosome is circularly permuted and terminally redundant. The DNA is often only "partially permuted" in that the ends are not completely randomly distributed across the whole genome sequence, but are found distributed over only a portion of the genome. This is the expected result if packaging series all start at the same place (a *pac* site), and terminal redundancy size and series lengths are limited; i.e., ends are all located in a region adjacent to the *pac* site for a distance that depends on the size of the terminal redundancy and on the number of events in packaging series (**Fig. 7.1**). Headful packaging phages do not have cohesive ends; their DNA ends are usually thought to be blunt due to their ability to be ligated to other blunt DNA ends *(24)*, but in fact the precise nature of their ends has been difficult to determine unambiguously, because of the many different end positions present in any DNA preparation.

A further complication to the analysis of headful packaging phage chromosomes is that during packaging the determination of when the capsid is full of DNA is imprecise, so somewhat different lengths of DNA are packaged in different individual virions (**Fig. 7.1**). This variation is about ±2% the genome length or ±700–1,000 bp in the few cases that have been studied *(23, 25, 26)*. Yet another complication is that the site of the initiation cleavage for headful packaging series is not precise in the headful phages analyzed to date, and alternative initiation cuts are scattered over regions that range from 9 bp in phage SPP1 to about 2,000 bp in phage Sf6 *(27, 28, 29, 30, 31, 32)*. The overall results of such packaging events, where both packaging initiation and chromosome length are imprecise, are DNA molecules from different individual virions whose end locations can lie at many if not all of the possible positions within a substantial region of the genome. We discuss below the diagnostic experimental features of headful packaging.

Since their virion chromosomes are not all the same length, one way of determining whether a phage utilizes a heedful-type packaging strategy is the greater width of the whole chromosome band relative to a similar-sized DNA molecule of precise length (e.g., phage λ virion DNA) after pulsed-field electrophoresis. This is not described in detail here, but see. for example. **Fig. 7.2c** in **ref**. *(31)* and **Fig. 7.2b** in **ref**. *(32)*. We note that phage Mu uses a headful-types packaging mechanism in spite of its very different packaging substrate (see below), so like the other headful packaging phages discussed in this section, it

also generates chromosomes of somewhat variable length *(33)*. Another experimental indicator of headful packaging is generalized transduction. Because of the lack of terminase sequence specificity in the headful cleavages, if packaging mistakenly initiates on host DNA, it is efficiently packaged into functional virion-like transducing particles (unlike in the COS phages, where sequence specificity is required at both the initiation and the headful cleavages and special experimental strategies are required to show generalized transduction *(34)*).

A more informative analysis of headful packaging can be obtained from analysis of the virion DNA's restriction fragment pattern. When terminally redundant, circularly permuted, headful packaged DNA is restricted, all of the true restriction fragments (fragments with restriction enzyme cleavages at *both* ends) that would be generated from a circular version of the genome sequence are typically present in at least some virion DNA molecules. Thus, the electrophoresis band pattern is *at minimum* that which *would have been* generated by circular DNA. If the series initiation cleavage near the *pac* site is relatively precise (all cleavages occur within a few hundred base pair regions), restriction of DNA from the first packaging event in any series will generate a discrete DNA fragment with a packaging initiation cleavage at its left end and a restriction cleavage at its right end (orientation as in **Fig. 7.1**). This series initiation-end fragment is called the "pac fragment" *(3)*, and it is present in fewer copies than the true restriction fragments, since it is only created from the first DNA packaging event of any packaging series (the molar ratio of the pac fragment to the true restriction fragments that are present in essentially all molecules correspond to the average packaging series length) *(3, 27, 35)*. Because of the substantial imprecision in the headful measurement of packaged DNA length (above), the packaging termination-generated end is imprecise. Thus, the right end restriction fragment from the first event in a packaging series and all terminal fragments (from both ends) generated by subsequent events in that series are variable in size and are so spread out in the gel background that they are nearly "invisible" in ethidium bromide-stained electrophoresis gels. Thus, *if* series initiation is relatively precise (as it is in phages P22, SPP1, and P1, for example), the restriction pattern of headful packaged DNA will consist of all the fragments expected from a circular genome *plus* a submolar pac fragment. The phage P22 restriction pattern is shown in **Fig. 7.2B**; *see* also **Fig. 7.1A** in *(3)*. When present, this type of restriction pattern is considered to be *diagnostic of headful packaging*. We note that the position of the terminase-generated end of the pac fragment is very near the *pac* site in the cases studied (summarized in **ref**. 19).

There are potential complications to such an analysis that must be understood. Of course, for any given restriction enzyme,

the pac fragment can be small and run off the gel or be obscured by a true restriction fragment, so the absence of an apparent pac fragment does not prove the absence of headful packaging, especially if no genome sequence is available (see also below); it may be necessary to try a number of different restriction enzymes to find ones that display the pac fragment unambiguously. This type of analysis can also be complicated somewhat if the circular permutation is sufficiently limited that all ends fall in a relatively small fraction of the genome; under these conditions, the true restriction fragments that extend across this region can also be present in submolar amounts or even missing altogether. This type of analysis is much more robust if a restriction map of the genome is available. Thus, several restriction patterns should agree on the location of the terminase-generated end of the pac fragment. Finally, certain special combinations of fragment size, terminal redundancy length, and headful precision can give rise to visible diffuse gel bands from the non-series initiation ends (*see* **Fig. 7.2a** in **ref**. 23).

Some headful terminases appear to be able to move substantial distances along the DNA between *pac* site recognition and cleavage of the DNA to initiate a packaging series. If it moves only a short distance, a discrete pac fragment is generated as discussed above. However, if it can move long distances, the series initiation cleavage by terminase becomes too imprecise to allow the generation of an easily visualized pac fragment in electrophoresis gels. This is the case for phages Sf6 and ES18 *(31, 32)* and perhaps T4 *(36)*. Then it too is an "invisible" diffuse band in a stained electrophoresis gel of restricted virion DNA, and the pattern of restriction fragments will simply be that expected from a circular genome; i.e., *all* the terminal fragments are so variable in length that they are lost in the background. Note in **Fig. 7.2** that these terminal fragments in headful phages P22 and Sf6 show as DNA staining in the background (between bands), and phages λ and SP6 which have unique ends have much less background straining material. In some fortuitous situations, a diffuse pac fragment can be seen as a faint, fuzzy-stained bands as in the MluI and BglII digested phage Sf6 DNAs shown in **Fig. 7.2C**. However, with this type of phage, Southern analysis using a probe that hybridizes to all of the variably sized pac fragments will specifically visualize the diffuse pac fragments band (and the true restriction fragment that includes the pac fragment) *(31, 32)*. Nucleotide sequence information is required in choosing the DNA probe. Where they have been studied, headful packaging phage *pac* sites usually lie within or near the small terminase subunit gene and packaging proceeds in the direction in which that gene is transcribed. As in the COS phages (above) in nearly all characterized headful packaging phages the large terminase subunit gene and portal protein genes are close and transcriptionally downstream

from the small subunit gene. Thus, a Southern probe from within the large terminase or portal gene will usually hybridize to any (diffuse or not) pac fragments that might initiate near a *pac* recognition site near the small terminase gene. One known exception is S*treptococcus* phage MM1, where the *pac* site is about 2 Kbp transcriptionally downstream of the large terminase gene, however, it is possible that this is a case of the small terminase gene, which is not recognizable in the MM1 sequence, being in a non-canonical location *(37)*. Identifying these diffuse bands generated by several different restriction enzymes effectively locates the variable end of the pac fragment and thus the region where packaging series begins.

Any method that can locate the chromosome ends can in theory elucidate the same things that the above restriction analysis does, and electron microscopic heteroduplex and partial denaturation mapping as well as cloning of terminal fragments have been used in this context. In fact, Tye et al. *(20)* used this type of electron microscopic analysis to be the first to deduce the "*pac* site—packaging series" strategy utilized by phage P22. More recently, Loessner et al. *(38)* have applied this technology to the analysis of *Listeria* phage A118. Plunkett et al. *(39)* deduced tentative chromosome end locations by analysis of the locations of DNA clones in a random library of phage 933W chromosomal DNA. Both methods are labor-intensive because of the need to manually examine many independent DNA molecules to gain the necessary statistical power. We will not describe these methods in detail here.

A final note of caution is warranted regarding interpretation of *apparent*, completely "random permutation" of phage chromosomes. Phages have not been studied to the point that we have cataloged all of their molecular lifestyles in detail, so packaging strategies other than those discussed here are possible. For example, the existence of headful packaging phages that initiate packaging, without a *pac* site, at genuinely random locations on the phage DNA remains possible. The observation of *apparently* completely randomly distributed virion DNA ends could be the result of the presence multiple *pac* sites, long packaging series, long terminal redundancy and/or terminase movements between recognition and DNA cleavage. On the other hand, it could be due to a "new" packaging strategy that somehow starts packaging randomly on the phage DNA without a *pac* site. The former explanations seem more likely since all phages that have been analyzed preferentially package their own DNA, and no truly random situation has been documented. The most relevant experimentally studied case is phage T4, which destroys all the non-T4 DNA in the infected cell and so *might* not need a nucleotide sequence target to recognize its own DNA. Its packaging strategy is not yet understood in every detail, but the current best working model is

that it too initiates packaging imprecisely at a recognition site in or near its small terminase gene *(36)*.

1.4 Short Exact Direct Repeat Ends

1.4.1 Best Studied Phages—T3 and T7

Phages of this type have direct double-stranded repeats at their termini which are a few hundred base pairs long and are exactly the same in every virion chromosome (i.e., they are not permuted). These chromosomes, when characterized, are thought to have blunt ends, again by the criterion of ligatability to other blunt ends. The terminal repeats are generated by a duplication of the direct repeat DNA in concert with packaging *(40, 41)* (**Fig. 7.1**). This type of DNA end structure could be overlooked when a phage genome sequence is determined by shotgun methods, since sequence assembly can merge the two ends to give a circular sequence. Analysis of appropriate restriction digests of this type of virion DNA will give rise to a gel pattern of fragments that has all equimolar fragments (and heating and cooling will not alter the pattern as with *cos* site phages) and terminal fragments which are not correctly predicted by such an artificially circularized sequence (except in the unlikely event of a fortuitous cleavage site in the short duplicated region; thus, multiple restriction digests should be analyzed in this way); **Fig. 7.2D** shows two restriction digests of T7-like phage SP6 DNA. In this type of phage, the approximate location of the ends can be determined by restriction mapping since the restriction maps of such molecules are linear, or if nucleotide sequence is available, approximate repeat locations can sometimes be predicted from the end locations in related phages. When the approximate locations of the chromosome ends are known, sequencing reactions initiated by primers that anneal to unique whole genome template sites internal to the terminal repeat and program synthesis across the repeats to the two ends can determine the exact end of the template strands by the position that synthesis stops and thus the length of the terminal repeats *(42, 43, 44)*.

1.5 Long Exact Direct Repeat Ends

1.5.1 Best Studied Phages—T5 and SPO1

Most "exact terminal repeat" phages that have been characterized have terminal repeats that have lengths in the one to a few hundred base pairs range as described in the previous section; however, several large, complex phages with genomes in the 130 kbp range, the best studied of which are *E. coli* phage T5 and *Bacillus subtilis* phage SPO1, have chromosomes with very long exact terminal direct repeats that are 10139 and 13185 bp, respectively (*(45)*; R. Hendrix, W. Huang, S. Casjens, G. Hatfull, M. Padulla and C. Stewart, unpublished results). The mechanism by which these long repeats are generated is not known, but as with the short exact repeat phages, there appears to be only one copy of the terminal repeat between genomes in replication-generated concatemers. This suggests that the repeat region is duplicated

prior to or during packaging. In these cases, shotgun sequencing methods will determine an *apparently* circular sequence. The long terminal redundancy can be noticed genetically *(46, 47)* or by electron microscopic analysis of heteroduplexes of nuclease-trimmed DNA *(48)*, but it is best discovered and studied by restriction mapping, which can locate the approximate position of the ends of such molecules on the sequence *(49, 50)*. Determination of the precise end sequences is more difficult since no uniquely templated primer can program a sequencing run all the way across the long terminal redundancy to the molecular end of a virion chromosome template. The best strategy is to isolate or clone terminal restriction fragments and use them to template sequencing reactions over the ends *(51)*.

1.6 Host DNA at Ends

1.6.1 Best Studied Phage—Mu

Phages like Mu, which replicate their genomes by duplicative transposition into host DNA, must package viral genomes that are randomly integrated into host DNA. The Mu terminase has not been studied in detail but it appears to recognize a *pac* site near one end of the genome and reach out from that point to initiate a packaging event by cutting in the adjacent host DNA *(52, 53)*. A headful type packaging event then extends from this initial cleavage, across the integrated phage DNA and beyond, to include about 1,800 bp of host DNA at the other side (**Fig. 7.1**) *(54)*. Again the ends are thought to be blunt by their ability to be ligated to other blunt ends (e.g., EcoRV ends in **ref**. *(55)*). Thus, each Mu DNA molecule has unique host DNA attached at both ends that is different in different virions. This terminal host DNA can be recognized "automatically" during a genome sequence determination project by the presence of these joined host sequences *(55, 56)*, but can be more difficult to recognize in the absence of this information. One straightforward way to detect is the presence of "frayed single-stranded ends" upon electron microscopic examination of heteroduplex molecules after strand separation and reannealing of virion DNA, since the host DNA at the ends has huge variety and strands with matching terminal host sequences almost never find one another during the reannealing. Details of this technique are not given here, but *see* **ref**. *(54)*.

1.7 Covalently Bound Terminal Proteins

1.7.1 Best Studied Phage—ϕ29

Bacillus subtilis phage ϕ29 and its relatives that infect other Gram-positive bacteria are the only tailed-phages currently known that have covalently bound proteins at the ends of their virion and replicating chromosomes. Such DNAs are typically recognized by the abnormally slow migration of their terminal DNA fragments in electrophoresis gels, and restoration to "normal" mobility by protease treatment. This is not discussed in detail here, but *see* **refs**. *(57, 58)*.

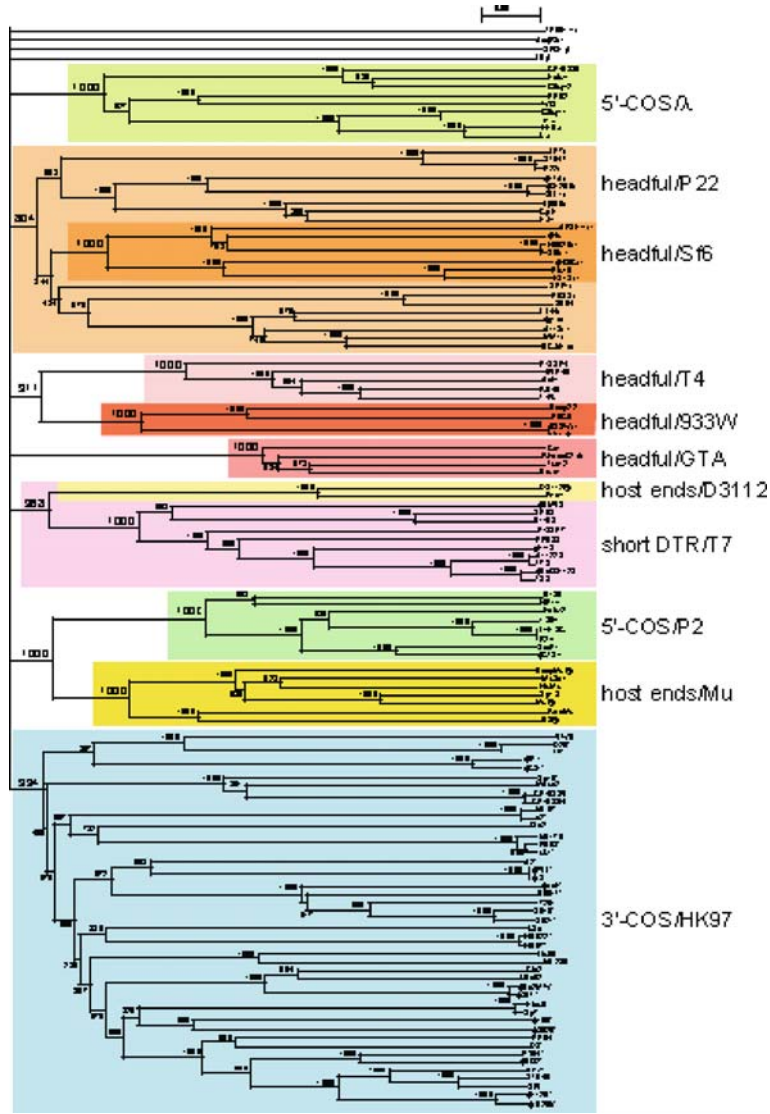


Fig. 7.3. Neighbor-joining tree of large terminase subunit amino acid sequences. A neighbor-joining tree of 123 tailed-phage terminase amino acid sequences was generated by CLUSTAL X *(59)*. The numbers near bifurcations are bootstrap values for 1,000 trials. Short bifurcating branches linking the major groups shown here were manually merged, since all had low bootstrap values, and the major deep branches are all shown as radiating from a single source. The names of the phages or prophages are shown at the right of each terminal branch (some prophage names are those proposed by Casjens *(13)*). Major robust, related groups of terminases are highlighted with *gray boxes*, and the packaging strategy and a prototype phage for each group is given at the far right. Those phages whose virion DNA termini structures have been experimentally determined are indicated as follows: #, 5′-cohesive ends in lambdoid phages; ^, 5′-cohesive ends in P2-like phages; *, 3′-cohesive ends; $, T7-like phages with direct terminal repeats and no circular permutation; , phages with long direct terminal repeats and no circular permutation; @, phages with host DNA at termini; †, headful packaging in non-T4-like phages, including P22 and the gene transfer agents (GTA); %, headful packaging in T4-like phages; •, headful packaging phages for which no

1.8 Prediction of Packaging Strategy and DNA End Structure from Terminase Amino Acid Sequence

If the amino acid sequence of a phage's large terminase subunit is known, the packaging strategy (and hence type of DNA ends) that a phage utilizes can often be successfully predicted from it. The terminase enzymes that create the virion DNA ends are quite varied, but still are among the most conserved tailed-phage proteins *(13)*. Comparative analysis among them has shown that they cluster according to the type of DNA end that they create *(32)*. Thus, if a terminase amino acid sequence falls *robustly* into one of the characterized clusters it very likely forms DNA ends that are similar to the other members of the group. **Fig. 7.3** shows such a tree where 3′-COS phages, two groups of 5′-COS phages, short exact direct terminal repeat phages, two apparent types of terminal host sequence phages, and *at least* five separable groups of headful packaging phages can be distinguished by their terminase amino acid sequences. It is also important to realize that the current picture of terminase diversity is not complete, and if a terminase sequence does not fall convincingly within a characterized group (currently including, for example, the long direct terminal repeat terminases of SPO1 and T5, headful terminases of P1, TP901-1, and Aaϕ23; COS phage VP16C, and short terminal repeat phage VpV260), the predictive value of such a comparison falls precipitously. In addition, as was pointed out by Casjens et al. *(32)*, the inclusion of some terminases such as those of 3′-COS phages TM4, MS1, and r1t in the tree shown in **Fig. 7.3** can lower some of the groups' bootstrap values. This type of prediction should of course not replace experimental analysis, but can be helpful in pointing the best way to proceed experimentally in a determination of tailed-phage chromosome end structure.

2 Materials

No unusual or special materials or techniques are required for the analysis of tailed-phage virion chromosomal restriction patterns in electrophoresis gels, and protocols for such analyses can be found in any molecular biology laboratory manual *(60)*.

Fig. 7.3. (continued) obvious pac fragment band has been identified in ethidium stained electrophoresis gels of restricted DNA (see text; the darker orange boxed "headful/Sf6" subgroup within the "headful/P22" group appear to be one branch of this type of terminase). This figure is modified from Fig. 7.6 of Casjens et al. *(32)*.

3 Methods and Practical Considerations

3.1 DNA Release and Isolation from Virions

1. A solution of CsCl equilibrium density or step gradient-purified virions is made in 0.25% sodium lauryl sulfate (SDS), 50 mM Tris–Cl (pH 8.0), and 25 mM ethylenediaminetetraacetic acid (EDTA) by addition of appropriate amounts from 20% SDS, 1 M Tris–Cl, 0.5 M EDTA stock solutions.
2. Incubate at 75–80 °C for 15 min.
3. Add potassium acetate to a final concentration of 0.625 M from a 2 M stock and mix well.
4. Chill on ice for 60 min.
5. Remove the potassium-SDS precipitate by centrifugation at 10,000 rpm in a microfuge for 15 min at 4 °C.
6. Add two volumes of ethanol and wind the released DNA out of solution on the tip of a sterile Pasteur pipet.
7. Rinse the DNA on the pipet tip by dipping it into room temperature 70% ethanol, air dry for a few minutes until no obvious liquid remains, and dissolve by dipping the tip into 200 μl of 10 mM Tris–Cl (pH 8.0), 1 mM EDTA. The partially hydrated DNA will release from the pipet tip within a few minutes; then let the DNA dissolve at 4 °C for at least 12 h before use. Over drying can make resuspension more difficult.
8. DNA prepared from purified phage in this way should be suitable for subsequent manipulations. If it appears to be degraded (as assayed by electrophoretic analysis) by contaminating nucleases upon incubation in Mg^{++} containing buffer, shake it gently with an equal volume of equilibrated phenol at room temperature for 10 min, ethanol precipitate, and resuspend in 10 mM Tris–Cl (pH 8.0), 1 mM EDTA.

3.2 Cohesive End DNA Analysis

1. Cleave 1 μg of purified virion DNA with a restriction enzyme of choice (one that results in the display the two end fragments, joined and separated, at uncrowded gel positions, as determined by trial-and-error or from analysis of a genome sequence).
2. Heat the reaction mix from step 1 to 75–80 °C for 15 min, and divide into two equal portions. Chill one rapidly by placing it quickly in wet ice, and cool the other to room temperature slowly. Useful slow cooling can be achieved by programming a PCR cycler to cool from 75 to 24 °C over 40 min or by simply placing the tube in a 75 °C or 80 °C metal heating block and letting the block cool to room temperature on the bench top.
3. Separate and display the resulting DNA bands in an agarose electrophoresis gel and visualize bands by ethidium bromide staining *(60)*.

4. If the phage has COS ends, two fragments will be visible in the quick chilled sample that are missing (or nearly so) in the slow cooled sample. In the slow cooled sample, these two fragments should be joined as a larger fragment whose molecular weight should be their sum (*see* **Fig. 7.2A**).

3.3 Headful DNA Analysis

3.3.1 P22 Headful Type—Discrete Pac Fragment

1. Cleave 1 μg of purified virion DNA with a restriction enzyme that results in the display the DNA pac fragment at an uncrowded gel position. Enzyme choice can be determined by trial-and-error by searching for enzymes that yield a single submolar DNA fragment or from analysis of a genome sequence and assuming that the *pac* site is within the small terminase gene, just upstream of the large terminase gene (the latter is not foolproof since there are known exceptions to this location, see above).
2. Separate and display the resulting DNA bands in an agarose electrophoresis gel and visualize bands by ethidium bromide staining *(60)*.
3. Since submolar DNA bands can be present due to partial digestion, excess restriction enzyme should be used, and several restriction digests that each display a single submolar band should be visualized before this type of headful packaging is considered to be convincing.
4. In order for this to be a robust conclusion, a restriction site map or genome sequence should be available to show that the putative *pac* cleavage-generated ends of the putative pac fragments lie at the same location in the genome for each of the different enzymes used.

3.3.2 Sf6 Type—Diffuse Pac Fragment that Is not Obviously Visible by Staining

1. Cleave 1 μg of purified virion DNA with a restriction enzyme that results in the display, the single submolar diffuse DNA pac fragment at an uncrowded gel position. Such an analysis in phages with Sf6-type type of packaging strategy is not recommended without some DNA sequence and other information on the phage under study; for example, one might know that the phage DNA is somewhat variable in length (above) or the terminase is highly related to known headful terminases, but efforts to visualize a pac fragment by staining have proven unsuccessful. Sequence information is useful because the choice of probe for the Southern analysis (below) requires at least a good estimate of where the pac fragments will lie on the genome and so how they will be displayed in the electrophoresis gel relative to the true restriction fragments. As with the P22-type phages (sharp pac fragment band) above, this "best guess" is that the *pac* site is within the small terminase gene and that packaging proceeds from there in the direction of the portal protein gene (again, the latter is not foolproof, since there are known exceptions, and the large

virulent phages are not yet "well-studied" in this regard, see above).

2. Separate and display the resulting DNA bands in an agarose electrophoresis gel *(60)*.
3. Transfer the DNA to a membrane and perform Southern analysis *(31, 60, 61)* with a suitable DNA probe. The DNA probe can be either a cloned or a PCR amplified fragment of the phage DNA that is chosen to hybridize to the pac fragments; typically the probe will be within the large terminase or portal gene. This probe will also hybridize to the true restriction fragment that covers the pac fragment and restriction enzymes and gel conditions should be chosen that results in good separation of these two gel bands.
4. As with the P22 "sharp-pac fragment band" type analysis above, for this to be a robust conclusion one should show by restriction site mapping that the putative *pac* cleavage-generated ends of the pac fragments lie at the same location in the genome for each of the different enzymes used.

4 Notes

As was mentioned above, the actual laboratory techniques used are standard ones that molecular biology laboratories will be familiar with. Thus, success or failure in the analysis of tailed-phage virion DNA end structure and packaging strategy is much more dependent upon the specific analysis strategy chosen than on technical process or protocol details. We have therefore emphasized strategic issues in our discussion. Perhaps the most important technical aspect in such an analysis is the use of DNA that is prepared from highly purified virions. Virions should whenever possible be purified by a method that takes advantage of their unusual buoyant density (between those of nucleic acids and protein), and cesium chloride gradient centrifugation has historically been the method of choice for purifying phage particles according to particle density. In general, true equilibrium sedimentation is not essential and considerable time can be saved by the use of CsCl "step gradients" as described in Earnshaw et al. *(62)*. There are a few exceptional tailed-phage virions (e.g., phage ES18 *(32)*) which are apparently impermeable to Cs^+ ions and so band in such gradients a position that does not separate them from the bulk of the proteinaceous cellular materials; then one is largely relegated to methods that separate on the basis of size such as differential and sucrose gradient centrifugation.

Acknowledgements

We thank Roger Hendrix for phage SF6 and Miriam Susskind for the essence of the protocol for DNA isolation from virions. The authors' research was supported by NSF grant MCB-990526 to SRC.

References

1. Mousset, S. & Thomas, R. (1969) Ter, a function which generates the ends of the mature lambda chromosome. *Nature* **221**, 242–244.
2. Feiss, M. & Campbell, A. (1974) Duplication of the bacteriophage lambda cohesive end site: genetic studies. *J. Mol. Biol.* **83**, 527–540.
3. Jackson, E. N., Jackson, D. A. & Deans, R. J. (1978) EcoRI analysis of bacteriophage P22 DNA packaging. *J. Mol. Biol.* **118**, 365–388.
4. Emmons, S. W. (1974) Bacteriophage lambda derivatives carrying two copies of the cohesive end site. *J. Mol. Biol.* **83**, 511–525.
5. Adams, M. B., Hayden, M. & Casjens, S. (1983) On the sequential packaging of bacteriophage P22 DNA. *J. Virol.* **46**, 673–677.
6. Simpson, A. A., Tao, Y., Leiman, P. G., Badasso, M. O., He, Y., Jardine, P. J., Olson, N. H., Morais, M. C., Grimes, S., Anderson, D. L., Baker, T. S. & Rossmann, M. G. (2000) Structure of the bacteriophage f29 DNA packaging motor. *Nature* **408**, 745–750.
7. Hershey, A. D. & Burgi, E. (1965) Complementary structure of interacting sites at the ends of lambda DNA molecules. *Proc. Natl. Acad. Sci. USA* **53**, 325–330.
8. Wu, R. & Taylor, E. (1971) Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage lambda DNA. *J. Mol. Biol.* **57**, 491–511.
9. Ellis, D. M. & Dean, D. H. (1985) Nucleotide sequence of the cohesive single-stranded ends of Bacillus subtilis temperate bacteriophage f105. *J. Virol.* **55**, 513–515.
10. Padmanabhan, R., Wu, R. & Calendar, R. (1974) Complete nucleotide sequence of the cohesive ends of bacteriophage P2 deoxyribonucleic acid. *J. Biol. Chem.* **249**, 6197–6207.
11. Fitzmaurice, W. P., Waldman, A. S., Benjamin, R. C., Huang, P. C. & Scocca, J. J. (1984) Nucleotide sequence and properties of the cohesive DNA termini from bacteriophage HP1c1 of *Haemophilus influenzae* Rd. *Gene* **31**, 197–203.
12. Juhala, R. J., Ford, M. E., Duda, R. L., Youlton, A., Hatfull, G. F. & Hendrix, R. W. (2000) Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. *J. Mol. Biol.* **299**, 27–51.
13. Casjens, S. (2003) Prophages in bacterial genomics: What have we learned so far? *Molec. Microbiol.* **249**, 277–300.
14. Hatfull, G. F. & Sarkis, G. J. (1993) DNA sequence, structure and gene expression of mycobacteriophage L5: a phage system for mycobacterial genetics. *Molec. Microbiol.* **7**, 395–405.
15. Ford, M. E., Sarkis, G. J., Belanger, A. E., Hendrix, R. W. & Hatfull, G. F. (1998) Genome structure of mycobacteriophage D29: implications for phage evolution. *J. Mol. Biol.* **279**, 143–164.
16. Lubbers, M., Ward, L., Beresford, T., Jarvis, B. & Jarvis, A. (1994) Sequencing and analysis of the cos region of the lactococcal bacteriophage c2. *Mol. Gen. Genet.* **245**, 160–166.
17. van Sinderen, D., Karsens, H., Kok, J., Terpstra, P., Ruiters, M. H., Venema, G. & Nauta, A. (1996) Sequence analysis and molecular characterization of the temperate lactococcal bacteriophage r1t. *Molec. Microbiol.* **19**, 1343–1355.
18. Streisinger, G., Enrich, J. & Stahl, M. (1967) Chromosome structure in T4. III. Terminal redundancy and length determination. *Proc. Natl'l. Acad. Sci., U.S.A.* **57**, 292–295.
19. Wu, H., Sampson, L., Parr, R. & Casjens, S. (2002) The DNA site utilized by bacteriophage P22 for initiation of DNA packaging. *Molec. Microbiol.* **45**, 1631–1646.
20. Tye, B. K., Huberman, J. A. & Botstein, D. (1974) Non-random circular permutation of phage P22 DNA. *J. Mol. Biol.* **85**, 501–528.
21. Moore, S. D. & Prevelige, P. E., Jr. (2002) Bacteriophage P22 portal vertex formation in vivo. *J. Mol. Biol.* **315**, 975–994.
22. Weigele, P. R., Sampson, L., Winn-Stapley, D. & Casjens, S. R. (2005) Molecular genetics of

bacteriophage P22 scaffolding protein's functional domains. *J. Mol. Biol.* **3**48, 831–844.
23. Casjens, S. & Hayden, M. (1988) Analysis in vivo of the bacteriophage P22 headful nuclease. *J. Mol. Biol.* **1**99, 467–474.
24. Schmieger, H., Taleghani, K. M., Meierl, A. & Weiss, L. (1990) A molecular analysis of terminase cuts in headful packaging of *Salmonella* phage P22. *Mol. Gen. Genet.* **2**21, 199–202.
25. Chow, L. T. & Bukhari, A. I. (1978) Heteroduplex electron microscopy of phage Mu mutants containing IS1 insertions and chloramphenicol resistance transposons. *Gene* **3**, 333–346.
26. Humphreys, G. O. & Trautner, T. A. (1981) Maturation of bacteriophage SPPI DNA: limited precision in the sizing of mature bacteriophage genomes. *J. Virol.* **3**7, 832–835.
27. Casjens, S. & Huang, W. M. (1982) Initiation of sequential packaging of bacteriophage P22 DNA. *J. Mol. Biol.* **1**57, 287–298.
28. Deichelbohrer, I., Alonso, J. C., Luder, G. & Trautner, T. A. (1985) Plasmid transduction by *Bacillus subtilis* bacteriophage SPP1: effects of DNA homology between plasmid and bacteriophage. *J. Bacteriol.* **1**62, 1238–1243.
29. Sternberg, N. & Coulby, J. (1987) Recognition and cleavage of the bacteriophage P1 packaging site (*pac*). II. Functional limits of pac and location of pac cleavage termini. *J. Mol. Biol.* **1**94, 469–479.
30. Casjens, S., Sampson, L., Randall, S., Eppler, K., Wu, H., Petri, J. B. & Schmieger, H. (1992) Molecular genetic analysis of bacteriophage P22 gene *3* product, a protein involved in the initiation of headful DNA packaging. *J. Mol. Biol.* **2**27, 1086–1099.
31. Casjens, S., Winn-Stapley, D., Gilcrease, E., Moreno, R., Kühlewein, C., Chua, J. E., Manning, P. A., Inwood, W. & Clark, A. J. (2004) The chromosome of *Shigella flexneri* bacteriophage Sf6: complete nucleotide sequence, genetic mosaicism, and DNA packaging. *J. Mol. Biol.* **3**39, 379–394.
32. Casjens, S. R., Gilcrease, E. B., Winn-Stapley, D. A., Schicklmaier, P., Schmieger, H., Pedulla, M. L., Ford, M. E., Houtz, J. M., Hatfull, G. F. & Hendrix, R. W. (2005) The generalized transducing *Salmonella* bacteriophage ES18: complete genome sequence and DNA packaging strategy. *J. Bacteriol.* **1**87, 1091–1104.
33. Chow, L. T. & Bukhari, A. I. (1977). Bacteriophage Mu genome: structural studies on Mu DNA and Mu mutants carrying insertions. In *DNA insertion elements, plasmids, and episomes* (Bukhari, A. I., Shapiro, J. A. & Adhya, S. L., eds.), pp. 295–306. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
34. Sternberg, N. (1986) The production of generalized transducing phage by bacteriophage lambda. *Gene* **5**0, 69–85.
35. Bachi, B. & Arber, W. (1977) Physical mapping of BglII, BamHI, EcoRI, HindIII and PstI restriction fragments of bacteriophage P1 DNA. *Mol. Gen. Genet.* **1**53, 311–324.
36. Lin, H. & Black, L. W. (1998) DNA requirements in vivo for phage T4 packaging. *Virology* **2**42, 118–127.
37. Obregon, V., Garcia, J. L., Garcia, E., Lopez, R. & Garcia, P. (2004) Peculiarities of the DNA of MM1, a temperate phage of *Streptococcus pneumoniae*. *Int. Microbiol.* 7, 133–137.
38. Loessner, M. J., Inman, R. B., Lauer, P. & Calendar, R. (2000) Complete nucleotide sequence, molecular analysis and genome structure of bacteriophage A118 of *Listeria monocytogenes*: implications for phage evolution. *Molec. Microbiol.* **3**5, 324–340.
39. Plunkett, G., 3rd, Rose, D. J., Durfee, T. J. & 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.* **1**81, 1767–1778.
40. Chung, Y. B., Nardone, C. & Hinkle, D. C. (1990) Bacteriophage T7 DNA packaging. III. A "hairpin" end formed on T7 concatemers may be an intermediate in the processing reaction. *J. Mol. Biol.* **2**16, 939–948.
41. Zhang, X. & Studier, F. W. (2004) Multiple roles of T7 RNA polymerase and T7 lysozyme during bacteriophage T7 infection. *J. Mol. Biol.* **3**40, 707–730.
42. Dunn, J. & Studier, W. (1983) Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. *J. Mol. Biol.* **1**66, 477–535.
43. Dobbins, A. T., George, M., Jr., Basham, D. A., Ford, M. E., Houtz, J. M., Pedulla, M. L., Lawrence, J. G., Hatfull, G. F. & Hendrix, R. W. (2004) Complete genomic sequence of the virulent *Salmonella* bacteriophage SP6. *J. Bacteriol.* **1**86, 1933–1944.
44. Scholl, D., Kieleczawa, J., Kemp, P., Rush, J., Richardson, C. C., Merril, C., Adhya, S. & Molineux, I. J. (2004) Genomic analysis of bacteriophages SP6 and K1-5, an estranged subgroup of the T7 supergroup. *J. Mol. Biol.* **3**35, 1151–1171.
45. Wang, J., Jiang, Y., Vincent, M., Sun, Y., Yu, H., Wang, J., Bao, Q., Kong, H. & Hu, S. (2005) Complete genome sequence of bacteriophage T5. *Virology* **3**32, 45–65.
46. Fischhoff, D., MacNeil, D. & Kleckner, N. (1976) Terminal redundancy heterozygotes involving the first-step-transfer region of the bacteriophage T5 chromosome. *Genetics* **8**2, 145–159.

47. Cregg, J. M. & Stewart, C. R. (1978) Terminal redundancy of "high frequency of recombination" markers of *Bacillus subtilis* phage SPO1. *Virology* **8**6, 530–541.
48. Rhoades, M. & Rhoades, E. A. (1972) Terminal repetition in the DNA of bacteriophage T5. *J. Mol. Biol.* **6**9, 187–200.
49. Perkus, M. E. & Shub, D. A. (1985) Mapping the genes in the terminal redundancy of bacteriophage SPO1 with restriction endonucleases. *J. Virol.* **5**6, 40–48.
50. Wiest, J. S. & McCorquodale, D. J. (1990) Characterization of pre-early genes in the terminal repetition of bacteriophage BF23 DNA by nucleotide sequencing and restriction mapping. *Virology* **1**77, 745–754.
51. Panganiban, A. T. & Whiteley, H. R. (1983) *Bacillus subtilis* RNAase III cleavage sites in phage SP82 early mRNA. *Cell* **3**3, 907–913.
52. George, M. & Bukhari, A. I. (1981) Heterogeneous host DNA attached to the left end of mature bacteriophage Mu DNA. *Nature* **2**92, 175–176.
53. Groenen, M. A. & van de Putte, P. (1985) Mapping of a site for packaging of bacteriophage Mu DNA. *Virology* **1**44, 520–522.
54. Bukhari, A. I. & Taylor, A. L. (1975) Influence of insertions on packaging of host sequences covalently linked to bacteriophage Mu DNA. *Proc. Natl. Acad. Sci., U S A* 72, 4399–4403.
55. Morgan, G., Hatfull, G., Casjens, S. & Hendrix, R. (2002) Bacteriophage Mu genome sequence: analysis and comparison with Mu-like prophages in *Haemophilus, Neisseria* and *Deinococcus. J. Mol. Biol.* **3**17, 337–359.
56. Summer, E. J., Gonzalez, C. F., Carlisle, T., Mebane, L. M., Cass, A. M., Savva, C. G., LiPuma, J. & Young, R. (2004) *Burkholderia cenocepacia* phage BcepMu and a family of Mu-like phages encoding potential pathogenesis factors. *J. Mol. Biol.* **3**40, 49–65.
57. Ito, J. (1978) Bacteriophage f29 terminal protein: its association with the 5′ termini of the f29 genome. *J. Virol.* **2**8, 895–904.
58. Salas, M., Mellado, R. P. & Vinuela, E. (1978) Characterization of a protein covalently linked to the 5′ termini of the DNA of *Bacillus subtilis* phage f29. *J. Mol. Biol.* **1**19, 269–291.
59. Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G. & Gibson, T. J. (1998) Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.* **2**3, 403–405.
60. Maniatis, T., Fritsch, E. & Sambrook, J. (1982). Molecular cloning A laboratory manual, pp. pp150–163. Cold Spring Harbor Labortory, Cold Spring Harbor, NY.
61. Southern, E. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* **9**8, 503–517.
62. Earnshaw, W., Casjens, S. & Harrison, S. (1976) Assembly of the head of bacteriophage P22, X-ray diffraction from heads, proheads and related structures. *J. Mol. Biol.* **1**04, 387–410.

Chapter 8

In silico Characterization of DNA Motifs with Particular Reference to Promoters and Terminators

Rob Lavigne, André Villegas, Andrew M. Kropinksi

Abstract

Knowledge of the regulatory elements contained within bacteriophage genomes forms the basis for understanding genomic expression and organization. The *in silico* prediction of promoter and terminator sequences in phage genomes is a first step towards this understanding. In this chapter, a number of programs and resources to identify regulatory elements are listed and discussed. Combining the available web-resources and literature data optimizes these predictions and can thus aid in a more directed experimental identification of these regulatory elements.

Key words: promoter prediction, terminator prediction, regulatory elements.

1 Introduction

Phage genome analysis focuses primarily on the identification and functional assignment of genes, while an *in silico* search for regulatory elements is often neglected or performed in a non-systematic manner. This is mainly due to the speculative nature of these searches and the lack of specific programs, tailored for phage genome analysis.

Characterization of specific bacteriophages, such as T7, T4, T3, and λ, has provided us with important insights into their genomic organization, expression, and mode of replication *(1–3)*. This research has led to the recognition of transcription and replication processes within the sequence through conserved sequences (e.g., promoter recognition sites), subtle DNA patterns (e.g., GC/AT-rich regions, DNA asymmetry), or in the occurrence of secondary structures (e.g., terminators).

Martha R. J. Clokie, Andrew M. Kropinski (eds.), *Bacteriophages: Methods and Protocols, Volume 2: Molecular and Applied Aspects, vol. 502*, © 2009 Humana Press, a part of Springer Science+Business Media
DOI 10.1007/978-1-60327-565-1_8 Springerprotocols.com

Transcription relies on the presence of either phage-encoded or host RNA polymerases, recognizing specific DNA motifs within the sequence of the viral DNA. Transcription regulation of several phages during host infection has been excellently reviewed *(4)*. Generally speaking, prokaryotic host promoters have typically conserved boxes (−10/ − 35 class promoters, extended −10 class promoters), while phage-specific, single subunit RNA polymerases usually recognize a single stretch of about 20–23 bp that may be written as consensus sequences often in the form of sequence logos, displaying significant residues, and subtle sequence patterns *(5, 6)*. Variations in the structure of these conserved sequences often reflect transcriptional regulation or local specificity in the DNA–ligand interaction *(4, 7)*. Hence, prediction of these conserved sequences lies in the identification of these conserved DNA motifs.

Just as the upstream intergenic regions contain promoters and operator sequences, the downstream sequences often contain motifs associated with transcriptional termination. In bacteria and their phages, the latter can be divided into rho factor-independent (RITT) and rho-dependent transcriptional termination (RDTT). In spite of an increasing understanding of the role of rho in factor-dependent transcriptional termination, the *in silico* prediction of RDTT sites has not been achieved *(8)*. Therefore, approximately 50% of terminations cannot be predicted based on sequence analysis.

In the case of RITT, termination can be successfully predicted since this type of transcriptional termination is associated with a definable structure. The following diagram illustrates the chief characteristics of a rho-independent terminator (**Fig. 8.1**). It possesses (a) a stem structure high in Gs and Cs, (b) a small

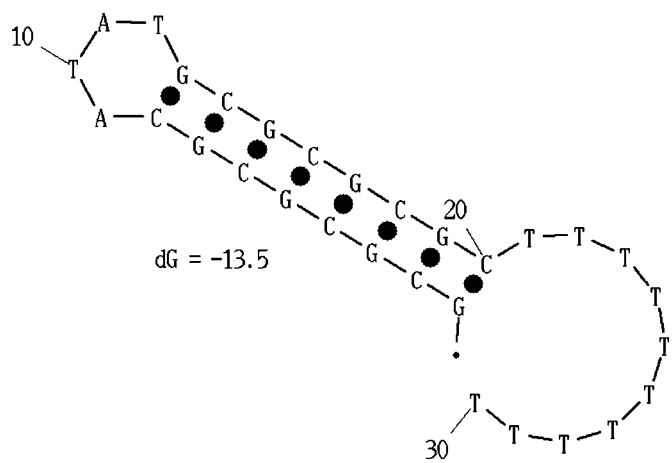


Fig. 8.1. Structure of a typical rho-independent transcriptional terminator as determined using Michael Zuker's program MFOLD.

loop which may contain a disproportionate number of UTGC in the case of T4 *(2)*, GAAA, and TTCG in *E. coli (9)* and TTT, AAT, TGA, or AAAA in *Bacillus subtilis (10)*; and, (c) a string of T/Us at the 3′ end.

In silico prediction of the origin of replication (*ori*) in phage genomes has been done based on DNA asymmetry in the genome. DNA asymmetry can be described as differences in nucleotide composition between complementary strands of DNA, due to different mutational pressure on the two strands. Both transcription and DNA replication can cause this asymmetry. Though often difficult to interpret, DNA asymmetry analyses have been performed to identify potential *ori* loci of T7-like phages *(11, 12)*. However, this approach of mapping the origin of replication by determining the number of strand-biased octamers at regular intervals is limited to phages transcribing genes in a single direction. The prediction program, developed by J.G. Lawrence, is to date not publicly available. A number of online programs (e.g., GraphDNA, GenSkew) make use of the base compositional skews to locate replication origins *(13,14,15)*.

2 Web-Based Resources

This section describes briefly some programs currently available on the World Wide Web, together with a brief description.

2.1 GraphDNA

DNA skew graphing is used for base compositional skew analysis with the tool provided at the Viral Bioinformatics Resource Center & Viral Bioinformatics, Canada. This well-documented site offers eight graphing options, including purine skew, DNA walker, AT, and GC skews.

http://athena.bioc.uvic.ca/workbench.php?tool=graphdna&db=

2.2 GenSkew

GenSkew provides double nucleotide skew analysis (e.g., AT, GC) with variable window and step sizes, and gives one the location of the replication origin.

http://mips.gsf.de/services/analysis/genskew/submit_html

2.3 Mfold

Mfold is an RNA and DNA folding package developed by Dr. Michael Zuker *(16)*. This program has an easy web-interface and presents the results in a wide variety of formats.
RNA: http://mfold.bioinfo.rpi.edu/cgi-bin/rna-form1.cgi
DNA: http://mfold.bioinfo.rpi.edu/cgi-bin/dna-form1.cgi

2.4 MEME/MAST

MEME/MAST motif discovery and search system *(17, 18)*. MEME allows you to predict conserved regions in a batch file of

related DNA strings (e.g., upstream regions of ORFs) and identifies motifs by comparison to sequence databases (MAST).

http://meme.sdsc.edu/meme/meme.html

2.5 CLUSTALW

CLUSTALW is a multiple sequence alignment program *(19)*. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities, and differences can be seen.

http://www.ebi.ac.uk/Tools/clustalw2/index.html

2.6 PHIRE

PHIRE searches bacteriophage genome sequences for conserved motifs *(20)*. Though aimed primarily at the identification of phage promoter sequences, conserved terminators or conserved repeat regions are readily identified. Though time-consuming, this program is theoretically suitable for all phages (**Fig. 8.2**).

http://www.biw.kuleuven.be/logt/PHIRE.htm

2.7 Ribex

Riboswitch Explorer *(21)* is a web tool for examining < 40 kb genomes for "riboswitch-like elements" which include tran-

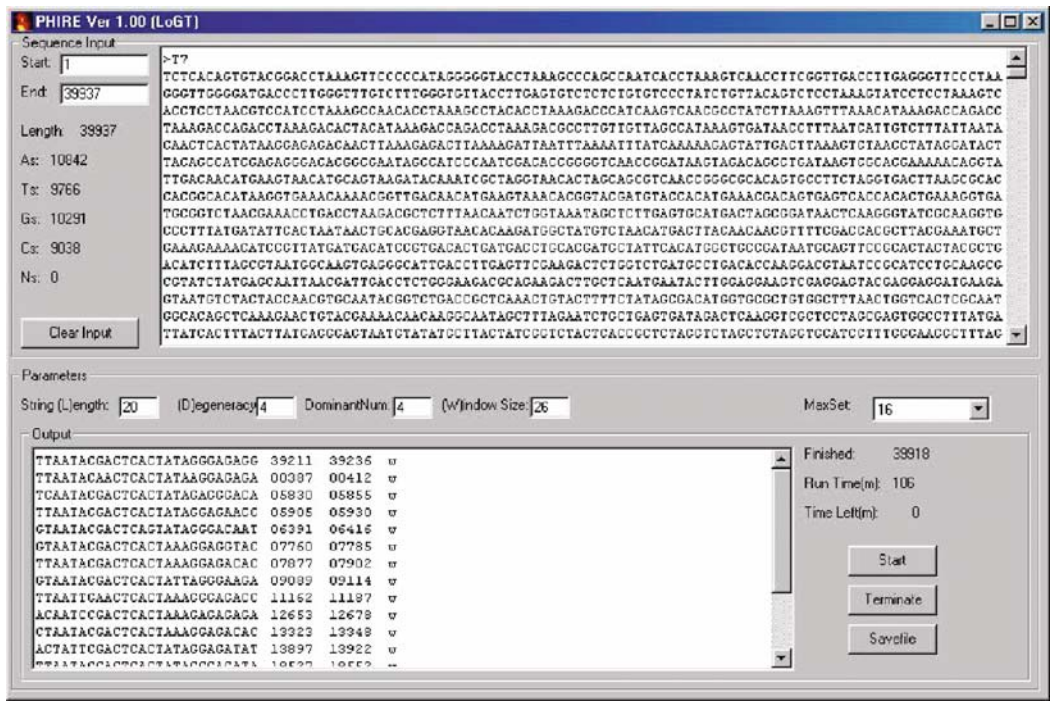


Fig. 8.2. PHIRE systematically compares all the DNA substrings of a specified length (L) to one another, allowing a limited number of mismatches (degeneracy D) to sort out and extract the largest sets (DominantNum) of substrings that represent a unique consensus. In this manner, the entire genome is analyzed on both the Watson and crick DNA strands. In order to visualize the sequences around the consensus sequence, the window size (W) can be adapted to include the sequences left and right of each selected DNA individual string.

scriptional attenuators (terminators, anti-terminators, and anti-antiterminators) and ORFs.

http://132.248.32.45:8080/cgi-bin/ribex.cgi

2.8 EMBOSS/PISE

Numerous implementations of the EMBOSS package are available on the www *(22)*. Using the Pise html-interface, these small programs are available for PC users. Einverted is one of many interesting programs offered, allowing genome wide identification of inverted repeats.

http://evol.biology.mcmaster.ca/EMBOSS-Pise.html

2.9 GeSTer

GeSTer is a Microsoft Windows-based program for determining stem-loop structures, including rho-independent terminators in annotated genomes. It is available from http://molbiol-tools.ca/Gester/. A new version of this useful software package is in development.

2.10 TransTerm

TransTermHP, an updated version of TransTerm is used for terminator prediction and is freely available in its UNIX format from the University of Maryland Center for Bioinformatics and Computational Biology *(23)*.

http://transterm.cbcb.umd.edu/?

2.11 Glimmer/RBSfinder

Glimmer/RBSfinder is a global ORF identification program and includes RBS prediction and terminator prediction (TransTerm) for improved ORF prediction *(24)*.

http://nbc11.biologie.uni-kl.de/framed/left/menu/auto/right/clusterinfo2/www/ (Open "Annotation" and choose Glimmer, RBSfinder with TransTerm)

2.12 RSAT

"Regulatory Sequence Analysis Tools" provide a series of modular computer programs specifically designed for the detection of regulatory signals in non-coding sequences *(25)*. Specifically, programs in this site may be useful for the identification of integration–host factor-binding sites in phage genomes.

http://rsat.ulb.ac.be/rsat/

2.13 BDGP: Neural Network Promoter Prediction

This program was developed for eukaryotic promoter prediction, but includes a prokaryote mode *(26)*.

http://www.fruitfly.org/seq_tools/promoter.html

2.14 PPP (Prokaryotic Promoter Prediction)

PPP developed by Aldert L. Zomer and Sacha A.F.T. van Hijum predicts promoter sequences and transcription factor-binding sites using a HMM model with hmmer.

http://bioinformatics.biol.rug.nl/websoftware/ppp/ppp_start.php

3 Implementation and Use

When searching for regulatory elements and DNA motifs, an important factor remains intuition and requires an "immersion" into the genome sequence, realizing that any predictions made should be rigorously scrutinized and formulated accordingly. Placing these predictions in a logical broader context (like genome organization or after comparative genomics with similar genomes) may increase their prediction probability. However, it should be added that less expected and unusual features should not be ignored. Any indications towards regulation can function as a potential "lead" for future experimental work and subsequently limit the time and financial effort required to obtain annotation.

3.1 Identification of Promoter Sequences

The availability of known promoter sequences from closely related phages allows searching specifically for these known promoter elements in the new genome sequence. The general sequence analysis package EMBOSS, combined with PISE (used to make this package available by generating a web-interface) contains an application called fuzznuc, which allows searching for known search patterns with a number of allowed mismatches *(22)*.

3.1.1 Promoters of Phage-Encoded RNA Polymerases

Nonetheless, most of the time the search for promoters implies looking for unknown, conserved sequences. Visual inspection of the intergenic regions or a ClustalW alignment *(19)* of ±150 bp regions upstream from open reading frames can reveal some of these conserved patterns *(27, 28)*. Also, other promoter prediction techniques have been developed, dependent on open reading frame predictions or on experimentally determined promoter sequences *(25, 29, 30)*.

An interesting example of this statistical approach is MEME/MAST (Multiple EM for Motif Elicitation/Motif Alignment and Search Tool) *(17, 18)*. Inputting the intergenic regions, or the upstream regions of each open reading frame, allows a versatile analysis of these sequences. For promoter sequences, zero or one occurrence of a motif is expected per sequence, while limiting the number of sites between 4 and 50 should cover the phage motifs. Furthermore, a motif length between 6 and 50 bp should cover these motifs.

A method developed by Chen and Schneider *(7)* uses information theory to predict T7-like promoters. In information theory, sequence motifs are analyzed based on available binding sites, allowing the content of specific nucleotides within the binding site to be evaluated *(31)*. Although this approach is sensitive compared to consensus analysis in which every position in the binding site is equivalent, the method relies heavily on the availability of known promoter sequences, requires new models

for different types of phages and is not available as a program package.

Recently, the PHIRE-program (Phage *In silico* Regulatory Elements) developed specifically to identify phage promoter sequences, relies solely on the availability of the genome sequence *(20)*. PHIRE compares all subsequences of the genome sequence to one another and extracts those subsequences most common within the genome. Because of the great number of executed comparisons, computation time is very high and increases exponentially with an increased genome size and is therefore limited to phage-size genomes.

PHIRE offers a deterministic approach to identify the promoter, but is limited to prior knowledge of the approximate length of the promoter sequence. Input of the genome sequence, the expected promoter length (L) and the allowed mismatches (D), allows generating a text file with the most dominant subsequences present in the genome (Dominantnum). After the analysis, the possibility to increase the window size (W) allows the flanking regions to be added to the output (a useful feature if the actual promoter length does not correspond completely to L).

PHIRE analyses on the currently available *Podoviridae* and *Myoviridae* genome sequences are summarized in **Tables 8.1** and **8.2**, respectively. Running PHIRE using its standard parameters allows extraction of potentially useful motifs in nearly all genomes. Interpretation of these identified motifs allows making functional predictions for many of them.

While PHIRE was developed especially for phage genomes, the use of MEME/MAST in the specific phage genomics domain has been very limited. Combination of MEME/MAST and PHIRE, both programs using different approaches, should maximize the generated output for prediction of the promoter sequences. Furthermore, these programs sometimes allow extraction of other features and motifs (e.g., conserved terminator sequences).

3.1.2 Promoters for Host RNA Polymerases

The search of bacterial promoters in phage sequences is not an easy task and journal reviewers quickly dismiss many predictions. For scanning of bacterial promoters, a number of programs exist (*See* **Section 2**). However, inputting phage genome sequences often leads to an overflow in obtained output. Hence, relevant predictions are difficult to distinguish from false positive results. These predictions are often made within context of the expected genome organization, while sometimes conservation within predicted promoters, observed using ClustalW, thus increasing prediction probability *(28)*.

3.2 Finding Transcriptional Terminators

Stem-loop structures and their secondary structure can be readily identified in DNA sequences using Accelrys's GCG StemLoop

Table 8.1
Summary of PHIRE predictions on all available *Podoviridae* genome sequences

NCBI Accession number	Name	Genome length (bp)	Identified PHIRE patterns	Predicted function
NC_000935	*Acyrthosiphon pisum* phage APSE-1	36,524	32,402–32,640	Region of short conserved repeats
NC_004165	*Bacillus* phage B103	18,630	14,650–14,685	Region of TAAAGA repeats
NC_002649	*Bacillus* phage GA-1	21,129	{CTATCTTTAGTATA}	Element of early promoters
NC_001423	*Bacillus* phage PZA	19,366	AATGTTTCACGTGGAACATT	Conserved inverted repeats
NC_007046	Bacteriophage 66	18,199	321–414	Short conserved repeats
NC_002730	Bacteriophage HK620	38,297	10,371–10,467	Ori locus
NC_006940	Bacteriophage KS7	40,794	{TAATAGWRYWCKATTAT}	
NC_003085	Bacteriophage Mx8	49,534	CSGCGGGCNTCGTCGTC	
NC_005045	Bacteriophage φKMV	42,519	CGACCCCTGCCCTACTCCGGCCTTAAA	Phage promoters
NC_001271	*Yersinia* phage φYeO3-12	39,600	AATTAACCCTCACTAAAGGGAG	Phage promoters
NC_003298	Bacteriophage T3	38,208	ATTAACCCTCACTAAAGGGAGA	Phage promoters
NC_005809	*Bordetella* phage BIP-1	42,638	33,235–33,686	Region of short conserved repeats
NC_005808	*Bordetella* phage BMP-1	42,663	33,231–33,658	Region of short conserved repeats
NC_005357	*Bordetella* phage BPP-1	42,493	33,240–33,547	Region of short conserved repeats
NC_005262	*Burkholderia cepacia* phage Bcep22	63,882	CGCGCGCCGCGCGCG	Conserved inverted repeats
			CGCGCTGGCGCRCGCGATC	

(continued)

Table 8.1 (continued)

NCBI Accession number	Name	Genome length (bp)	Identified PHIRE patterns	Predicted function
NC_003390	Cyanophage P60	47,872	37,829–38,382	Region of short conserved repeats
NC_006882	Cyanophage P-SSP7	44,970	AAAATTCTTCAAGTNTTA	
NC_004775	Enterobacteria phage ε15	39,671	ATTACCNAAANGGTAATW	Lysogeny-related element
NC_008152	Enterobacteria phage K1-5	44,385	ATTTACYNGACACTATAGAAGR	Phage promoters
NC_007456	Enterobacteria phage K1F	39,704	MCTAAACTATCACTNTAGGR	Phage promoters
NC_002371	Enterobacteria phage P22	41,724	32,859–32,925	Ori locus
NC_005344	Enterobacteria phage Sf6	39,043	28,387–28,483	Ori locus
NC_004831	Enterobacteria phage SP6	43,769	TTTANGKGACACTATAGRW	Phage promoters
NC_001604	Enterobacteria phage T7	39,937	ATACGACTCACTATAGGGAG/175–340	Phage promoters
NC_002515	*Mycoplasma* phage P1	11,660	AAASAAATTAA and 8273–8427	Conserved string and region
NC_007804	Enterobacteria phage φV10 virus	39,104	33,628–33,751	Ori locus
NC_005884	*Pseudomonas aeruginosa* phage PaP2	43,783	36,018–36,069	Region of short conserved repeats
NC_006552	*Pseudomonas aeruginosa* phage F116	65,195	19,478–19,647	“”
NC_004466	*Pseudomonas aeruginosa* phage PaP3	45,503	37,673–37,692	“”
NC_004665	*Pseudomonas* phage gh-1	37,359	TTAAAAACCCTCACTATGGC	Phage promoters
NC_002519	Roseophage SIO1	39,898	07,133–07,488	Replication region

(continued)

Table 8.1 (continued)

NCBI Accession number	Name	Genome length (bp)	Identified PHIRE patterns	Predicted function
NC_006949	*Salmonella* Typhimurium phage ES18	46,900	34,796–35,085	Region of short conserved repeats
NC_005841	*Salmonella* Typhimurium phage ST104	41,391	9,861–10,165	“”
NC_004348	*Salmonella* Typhimurium phage ST64T	40,679	11,347–11,561	“”
NC_004313	*Salmonella* Typhimurium phage ST64B	40,149	31,289–31,374	“”
NC_004679	*Staphylococcus aureus* phage φP68	18,227	17,794–17,891	“”
NC_004678	*Staphylococcus* phage 44AHJD	16,784	16,148–16,452	“”
NC_004814	*Streptococcus* phage C1	16,687	AGAAAATAATTTTAAAATTTT-20 bp-TAATACATAAATAAGAAAGA	
NC_001825	*Streptococcus* phage Cp-1	19,343	19,039–19,162/192–305	Conserved AT-rich regions
NC_005879	Vibriophage VP2	39,853	10,821–11,226	Conserved GC-rich repeat region
NC_007149	Vibriophage VP4	39,503	AATTAACCCTGACTATAGGAA	Phage promoters
NC_005891	Vibriophage VP5	39,786	SKGTGTCGACRTCGACACM	Phage promoters
NC_003907	Vibriophage VpV262	46,012	TCACCTGCTGTGATGTAC	
NC_004777	*Yersinia pestis* phage φA1122	37,555	TAATACGACTCACTAWAGGRRR	Phage promoters
AM265638	Bacteriophage LKD16	43,200	CGACCCCTGCCCTACTCCGGCCTTAAA	Phage promoters
AM265639	Bacteriophage LKA1	41,593	CGTAACCGCTGCACTCGCAG	Phage promoters

Table 8.2
Summary of PHIRE predictions on all available *Myoviridae* genome sequences

Accession number	Name	Genome length (bp)	Identified PHIRE patterns	Predicted function
NC_008208	*Aeromonas salmonicida* bacteriophage 25	161,475	YAAAAAWGGCCTCCGAAGAG SCCWWW	Phage promoters
NC_006884	Cyanophage P-SSM4	178,249	GGTAATGTCACAGGTAATCT repeats in 99,212–70,023	
NC_000866	Enterobacteria phage T4	168,903	TTTCACAAARYTGTTTACAA	Phage promoters
NC_006565	*Lactobacillus plantarum* phage LP65	131,522	AKKCTTTTTATATASAANA	Phage promoters
NC_005083	*Vibrio* phage KVP40	244,834	AAGAGAGAAMATT**ATG**	Conserved RBS sequence
NC_006883	Cyanophage P-SSM2	252,401	21,474–28,291	Region of short conserved repeats
NC_004687	*Mycobacterium* phage Bxz1	156,102	CTCGACGCCGACGTCGACGAG	Conserved inverted repeats
NC_007623	*Pseudomonas* phage φEL	211,215	WTTTYAAACCTACATTATY	Phage promoters
NC_007610	*Listeria* bacteriophage P100	131,384	92,184–92,278	Repeat region
			AANAAWGACAANAAAGAA	
NC_007066	Bacteriophage G1	138,715	95,189–95,541	Repeat region
NC_007023	Enterobacteria phage RB43	180,500	AAAAGGGGCGAAAGCCCCTTT	Conserved inverted repeats
NC_007022	*Aeromonas* phage 31	172,963	TTNAAAAAACAGTTTACAAYG	Phage promoters
NC_007021	*Staphylococcus* phage Twort	130,706	AAAAAAAGAAAAAGAAG	A-rich stretches
NC_006820	Bacteriophage S-PM2	196,280	TGSTGGTGNTGGTGGT	
NC_005880	*Staphylococcus* phage K	127,395	ATAAAAAAGWWAAAAAAGAA	Putative phage promoters

(continued)

Table 8.2 (continued)

Accession number	Name	Genome length (bp)	Identified PHIRE patterns	Predicted function
NC_005856	Enterobacteria phage P1	94,800	GCTCTAATAAAT	Conserved sigma 70-like promoter
NC_005260	Bacteriophage Aeh1	233,234	T-rich-CATGATGTATTCCTCAG	
NC_005135	Bacteriophage 44RR2.8t	173,591	TTATTATAGTCCCATCAAATC	Phage promoters
NC_005066	Enterobacteria phage RB49	164,018	TGAGGATTAGATT**ATG**	Conserved RBS sequence
NC_004928	Enterobacteria phage RB69	167,560	GATAGGTCTATAATAACATA	Phage promoters
NC_004735	Bacteriophage RM 378	129,908	TTTTTATTTAAATAAAAAGA	Intergenic AT-rich palindromes
NC_004629	*Pseudomonas* phage φKZ	280,334	ATGCCCTCCCTTCGGGGAGGGCTT	Phage promoters
			ATGCCCTCCCTTCGGGGAGGGCTT	Conserved inverted repeats
NC_004084	Virus φCh1	58,498	CGACGTCCAGATCGACGACG	Conserved inverted repeats
NC_005056	Bacteriophage Wφ	32,684	GCTGGTGATGCCGGTGCGCTGG	
NC_003444	*Shigella flexneri* phage V	37,074	28,654–28,678	Region of short conserved repeats
NC_003315	*Haemophilus* phage HP2	31,508	CAAACAAAGCAAAAAC	
NC_003313	Bacteriophage K139	33,106	AGAACTGRTTGARGTGAT	
NC_003278	Bacteriophage φCTX	35,580	ATGGCCGGCCTRR	
NC_000929	Enterobacteria phage Mu	36,717	TTTCCGACATGGAAA	Minor inverted repeats
NC_001895	Enterobacteria phage P2	33,593	GCCGGTGCGCTGGCGC	
NC_001697	*Haemophilus* phage HP1	32,355	AAAAAATAAAAGAAAATAA	
NC_001609	Bacteriophage P4	11,624	CATTTAAAGCCACTTAAAGC	

(continued)

Table 8.2 (continued)

Accession number	Name	Genome length (bp)	Identified PHIRE patterns	Predicted function
NC_008201	Bacteriophage φMhaA1-PHL101	34,525	18,294–18,357	AAGC-repeats
NC_008193	Bacteriophage F108	30,505	AAAAAAGANNAATAAAAAAG	
NC_001317	Enterobacteria phage 186	30,624	CRACATAAGTYCCATYAGGGGC repeat in 21,333–21,392 region	
NC_007917	*Clostridium difficile* phage φCD119	53,325	AAAAAATAAAATAAATAAA	
NC_005886	*Burkholderia cenocepacia* phage BcepB1A	47,399	36,215–36,410	SGGC repeat region
NC_005882	*Burkholderia cenocepacia* phage BcepMu	36,748	CGCCGCGCGCGCGGCCGCC	
NC_005342	*Burkholderia cepacia* phage Bcep43	48,024	CGCCGGCGACGGCCGC	
NC_005340	Bacteriophage PSP3	30,636	CATCGACCAGACTGTCG	
NC_005294	Bacteriophage EJ-1 provirus	42,935	4,390–4,430	AAAAAAAT repeats
NC_005263	*Burkholderia cenocepacia* phage Bcep1	48,177	CGCGCGCGGCGCCGGC	
NC_004827	Bacteriophage Aaφ23	43,033	AAAATAAGACCGCACTTT	Intergenic conserved motif
NC_004745	Bacteriophage L-413C	30,728	TGGTCATGTTGCTGGCGCTG	
NC_004456	*Vibrio harveyi* phage VHML	43,198	CCGTTTATTTGTTGGTTTGT	Intergenic conserved motif
NC_004333	*Burkholderia cepacia* phage Bcep781	48,247	CAGCGTGCCGGCGGCGGCCG	

and MFOLD programs (Accelrys Software Inc., San Diego, CA). As less expensive options, inexpensive software programs such as DNAMAN (Lynnon Corp., Vaudreuil-Dorion, QC, Canada) or online tools can be employed. The latter include EMBOSS's program EINVERTED *(32)*. The two approaches that can be employed are to examine the DNA sequence for inverted repeats and then see whether any of these have adjacent downstream T-rich regions. Alternatively, one can search for polyA or polyT regions and then screen the adjacent sequence for stem-loop structures using MFOLD.

One of the most useful software packages for specifically examining genome sequences for termination signals is GeSTer *(33)*. This has been used to analyze terminators in bacteriophage ε15 *(34), Bacillus subtilis, Escherichia coli, Mycobacterium, Neisseria, Xylella (33)*, and *Staphylococcus (35)*. GeSTer has been used on a laptop equipped with Microsoft Windows XP. Unfortunately, this program has problems dealing with the format of recent GenBank flatfiles (*.gbk). Replacing the term "locus_tag" with "gene," and deleting superfluous information can circumvent this. A minimal gbk file, derived from that of *E. coli* K12, is shown in **Fig. 8.3**.

In **Fig. 8.4** we see a screenshot of the opening menu, in which the "Max. Distance from ORF" has been adjusted from the default of 270 nt to a more realistic value of 100. On choosing the appropriate file and output file name ("Result" is the default name), the program will generate the data shown in **Fig. 8.4**.

This program generates a number of *.dat files whose names are logical and which can be opened using Notepad. I have chosen to view one of the palindromic sequences with the greatest ΔG value present on the regular (versus complementary) strand from the greatestdgreg.dat file (**Fig. 8.5**).

```
LOCUS       A00000     4639221 bp
FEATURES             Location/Qualifiers
     source          1..4639221
     CDS             190..255
                     /gene="thrL"
     CDS             337..2799
                     /gene="thrA"
     CDS             complement(5683..6459)
                     /gene="yaaA"
     CDS             complement(6529..7959)
                     /gene="yaaJ"
ORIGIN
        1 agcttttcat tctgactgca acgggcaata tgtctctgtg tggattaaaa aaagagtgtc
       61 tgatagcagc ttctgaactg gttacctgcc gtgagtaaat taaaatttta ttgacttagg

  4639141 gacacggcaa tgttgcaccg tttgctgcat gatattgaaa aaaatatcac caaataaaaa
  4639201 acgccttagt aagtattttt c
//
```

Fig. 8.3. A GenBank flatfile (*.gbk) formatted in this manner is easily read by GeSTer. The format can be modified using a text editor such as Notepad.

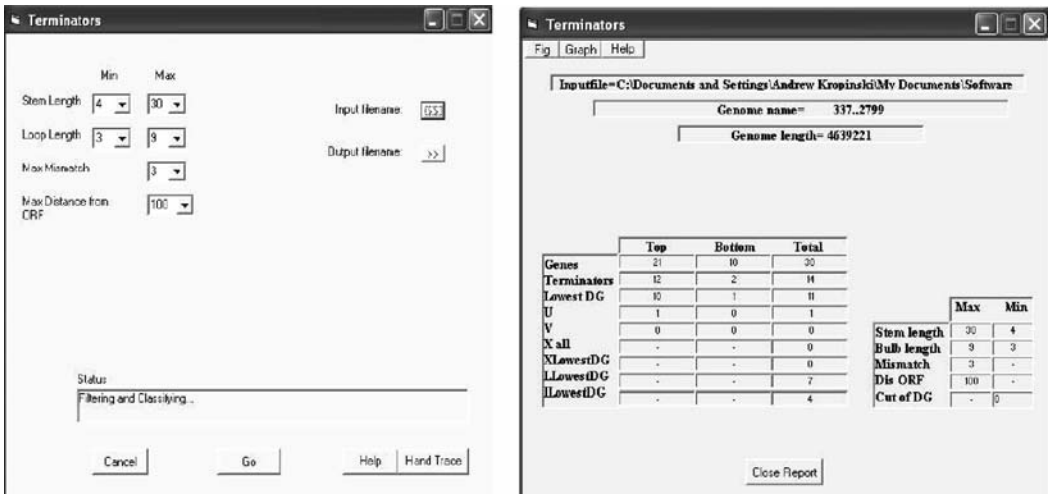


Fig. 8.4. The opening screen in GeSTer (*left*) and the results menu (*right).*

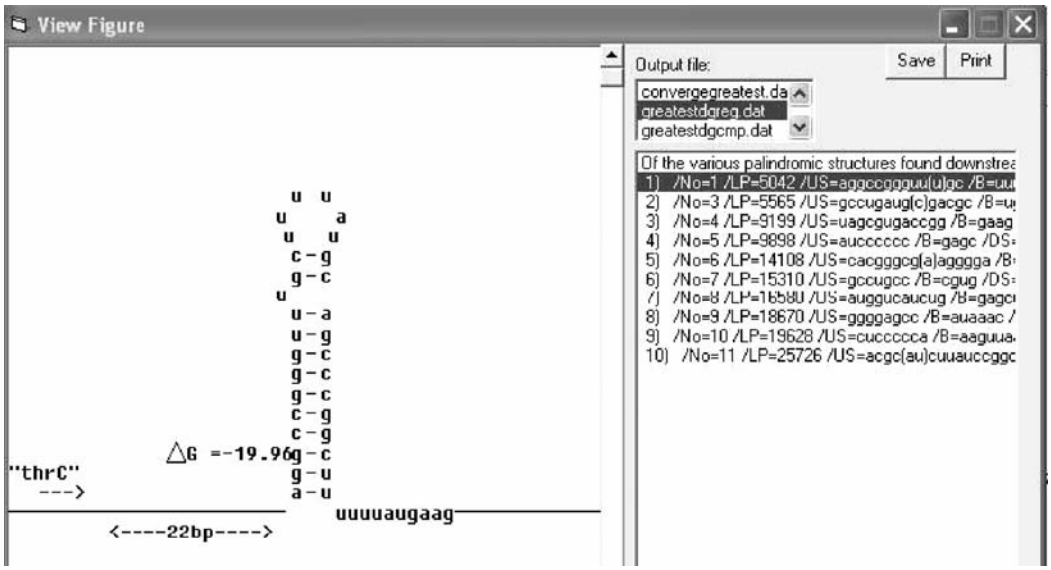


Fig. 8.5. Please note that you can access the numerous *.dat files that this program generates and see all the data on each terminator. The full data on terminator 1 is:/No=1/LP=5042/US=aggccggguu(u)gc/B=uuuuau/DS=gcagcccggcuu/T=uuuuaugaag/USL=13/DSL=12/SL=12/BL =6/Mm=0/Gp=1/DG= −19.96/G="thrC"/G>=3734/G<=5020/DS>=22/DS<=53/DM=37.5. The relevant information is US (upstream stem sequence), B (loop or bulb sequence), DS (downstream stem sequence), T (sequence of the tail, i.e., 3′ sequence), G (name of upstream gene), and DG (stability).

Another program is TransTerm, which was originally developed by the bioinformatics team at the Institute for Genomic Research. This program is also available at Michael Nuhn's online site packaged with two other TIGR programs Glimmer *(24)* and RBSfinder at http://nbc11.biologie.uni-kl.de/framed/left/menu/auto/right/clusterinfo2/www/. Run

in this mode it may miss a number of terminators. If you have the sequence of the DNA in GenBank format, this site also allows you to search downstream of all annotated genes.

Acknowledgements

R. Lavigne is a postdoctoral scientist of the Flemish FWO (Fonds voor Wetenschappelijk Onderzoek-Vlaanderen). A. Kropinski is supported by grants from the Public Health Agency of Canada's Laboratory for Foodborne Zoonoses and the Natural Sciences and Engineering Research Council of Canada.

References

1. Dunn, J. J. and Studier, F. W. (1983). Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. *J. Mol. Biol.*, **166**, 477–535.
2. Miller, E. S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T., and Ruger, W. (2003). Bacteriophage T4 genome. *Microbiol. Mol. Biol. Rev.*, **67**, 86–156.
3. Pajunen, M.I., Elizondo, M.R., Skurnik, M., Kieleczawa, J., Molineux, I.J. (2002). Complete nucleotide sequence and likely recombinatorial origin of bacteriophage T3. *J. Mol. Biol.*, **319**, 1115–1132.
4. Nechaev, S. and Severinov, K. (2003) Bacteriophage-induced modifications of host RNA polymerase. *Annu. Rev. Microbiol.*, **57**, 301–22.
5. Schneider, T.D. and Stephens, R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
6. Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: A sequence logo generator. *Genome Res.*, **14**, 1188–1190.
7. Chen, Z. and Schneider, T.D. (2005) Information theory based T7-like promoter models: classification of bacteriophages and differential evolution of promoters and their polymerases. *Nucleic Acids Res.*, **33**, 6172–6187.
8. Skordalakes, E. and Berger., J. M. (2003) Structure of the rho transcription terminator: mechanism of mRNA recognition and helicase loading. *Cell*, **114**, 135–146.
9. d'Aubenton-Carafa, Y., Brody, Y. and C. Thermes. (1990) Prediction of rho-independent *Escherichia coli* transcription terminators. A statistical analysis of their RNA stem-loop structures. *J. Mol. Biol.*, **216**, 835–858.
10. de Hoon, M. J. L., Makita, Y., Nakai, K. and S. Miyano (2005) Prediction of transcriptional terminators in *Bacillus subtilis* and related species. *PLoS Comp. Biol.*, **1**, 0212–0221.
11. Dobbins, A.T., George, M. Jr., Basham, D.A., Ford, M.E., Houtz, J.M., Pedulla, M.L., Lawrence, J.G., Hatfull, G.F. and Hendrix, R.W. (2004) Complete genomic sequence of the *Salmonella* bacteriophage SP6. *J. Bacteriol.*, **186**, 1933–1944.
12. Ceyssens, P-J., Lavigne, R., Chibeu, A., Mattheus, W., Hertveldt, K., Robben, J. and Volckaert, G (2006) Genomic analysis of *Pseudomonas aeruginosa* phages LKD16 and LKA1: Establishment of the ϕKMV subgroup within the T7 supergroup. *J. Bacteriol.*, in press.
13. Lobry, J.R. (1996) A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. *Biochimie* **78**, 323–326.
14. Lobry, J.R. (1999) Genomic landscapes. *Microbiol. Today* **26**, 16–164.
15. Grigoriev, A. (1998) Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res.* **26**, 2286–2290.
16. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
17. Bailey, T.L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
18. Bailey, T.L. and Gribskov, M. (1998) Methods and statistics for combining motif match scores. *J. Comput. Biol.* **5**, 211–221.
19. Thompson J.D., Higgins D.G. and Gibson T.J.(1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence

alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
20. Lavigne, R., Sun, W.D. and Volckaert, G. (2004) PHIRE, a deterministic approach to reveal regulatory elements in bacteriophage genomes. *Bioinformatics*, **20**, 629–635.
21. Abreu-Goodger, C. and Merino, E. (2005) RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements. *Nucleic Acids Res.*, **33**, W690–W692.
22. Rice, P., Longden, I. and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*, **16**, 276–277.
23. Ermolaeva, M. D., Khalak, H. G., White, O., Smith, H. O. and Salzberg, S. L. (2000). Prediction of transcription terminators in bacterial genomes. *J. Mol. Biol.*, **301**, 27–33.
24. Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.
25. van Helden, J., André, B. and Collado-Vides, J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281**, 827—842.
26. Reese, M.G. (2001) Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome, *Comput Chem.*, **26**, 51–6.
27. Mesyanzhinov, V.V., Robben, J., Grymonprez, B., Kostyuchenko, V.A., Burkal'tseva, M.V., Sykilinda, N.N., Krylov, V.N. and Volckaert, G. (2002) The genome of bacteriophage phiKZ of *Pseudomonas aeruginosa*. *J. Mol. Biol.*, **317**, 1–19.
28. Lavigne, R., Burkal'tseva, M.V., Robben, J., Sykilinda, N.N., Kurochkina, L.P., Grymonprez, B., Jonckx B., Krylov, V.N., Mesyanzhinov, V.V. and Volckaert, G. (2003) The genome of bacteriophage ϕKMV, a T7-like virus infecting *Pseudomonas aeruginosa*. *Virology*, **312**, 49–59.
29. Chen, Q.K., Hertz, G.Z. and Stormo, G.D. (1997) PromFD 1.0: a computer program that predicts eukaryotic pol II promoters using strings and IMD matrices. *Comput. Appl. Biosci.*, **13**, 29–35.
30. Marsan, L. and Sagot, M.F. (2000) Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *J. Comput. Biol.*, **7**, 345–62.
31. Schneider, T.D. (1996) Reading of DNA sequence logos: prediction of major groove binding by information theory. *Methods Enzymol.*, **274**, 445–455.
32. Mullan, L. J. and Bleasby, A. J. (2002) Short EMBOSS User Guide. European Molecular Biology Open Software Suite. *Brief. Bioinform.*, **3**, 92–94.
33. Unniraman, S., Prakash, R. and Nagaraja., V. (2002) Conserved economics of transcription termination in eubacteria. *Nucleic Acids Res.*, **30**, 675–684.
34. Kropinski, A. M., Kovalyova, I. V., Billington, S. J., Butts, B. D., Patrick, A. N., Guichard, J. A., Hutson, S. M., Sydlaske, A. D., Day, K. R., Falk, D. R. and McConnell, M. R. The genome of ε15, a serotype-converting, Group E1 *Salmonella* enterica-specific bacteriophage. (manuscript in preparation).
35. Wang, L., Trawick, J. D., Yamamoto, R. and Zamudio, C. (2004). Genome-wide operon prediction in *Staphylococcus aureus*. *Nucleic Acids Res.* **32**, 3689–3702.

Chapter 9

Molecular Phylogenetics: Testing Evolutionary Hypotheses

David A. Walsh and Adrian K. Sharma

Abstract

A common approach for investigating evolutionary relationships between genes and organisms is to compare extant DNA or protein sequences and infer an evolutionary tree. This methodology is known as molecular phylogenetics and may be the most informative means for exploring phage evolution, since there are few morphological features that can be used to differentiate between these tiny biological entities. In addition, phage genomes can be mosaic, meaning different genes or genomic regions can exhibit conflicting evolutionary histories due to lateral gene transfer or homologous recombination between different phage genomes. Molecular phylogenetics can be used to identify and study such genome mosaicism. This chapter provides a general introduction to the theory and methodology used to reconstruct phylogenetic relationships from molecular data. Also included is a discussion on how the evolutionary history of different genes within the same set of genomes can be compared, using a collection of T4-type phage genomes as an example. A compilation of programs and packages that are available for conducting phylogenetic analyses is supplied as an accompanying appendix.

Key words: Evolution, lateral gene transfer, phage mosaicism, maximum likelihood.

1 Introduction

Although phages may not be considered living organisms, the study of phage evolution is important as they are one of the most diverse and numerous biological entities on the planet *(1)*. Due to their small size and limited number of morphological features, the most informative method for investigating phage evolution is through the comparison of their molecular sequences. Such comparisons have been hampered in the past due to the lack of phage genomic sequences. While the number of phage genome sequences that exist in public databases is still relatively modest, the number is increasing exponentially. In addition,

Martha R. J. Clokie, Andrew M. Kropinski (eds.), *Bacteriophages: Methods and Protocols, Volume 2: Molecular and Applied Aspects, vol. 502*, © 2009 Humana Press, a part of Springer Science+Business Media
DOI 10.1007/978-1-60327-565-1_9 Springerprotocols.com

viral metagenome sequencing projects are generating a wealth of sequence data that is valuable for studying phage diversity and evolution *(2)*.

The increasing availability of phage sequences indicates molecular phylogenetics will be an important approach to the study of phage biology, particularly in reconstructing the evolutionary relationships between extant phage genomes. Phylogenetics has already been employed to uncover relationships between several phage types, for example, the T4-type bacteriophages *(3)*. Phylogenetic analysis of major capsid protein (g23) sequences recovered directly from the marine environment by PCR lead to the discovery of novel, geographically widespread T4-type phages *(4,5)*. In addition, phylogenetic analysis has challenged the official phage taxonomy based on morphology and genome organization *(6)* and whole genome analysis has lead to a proposed phage taxonomy based on overall genome similarity *(7)*.

The extent to which phage taxonomy based on whole genome sequence similarity represents a natural classification system or accurately reflects phage evolutionary history is presently unclear. Genome mosaicism due to gene exchange either by lateral gene transfer (LGT) *(8, 9)* or by homologous recombination *(10–12)* has emerged as a general characteristic of many prokaryotes. As in prokaryotes, modes of vertical and lateral evolution exist for phage and their genomes have been extensively described as modular *(13–15)*. In addition, genes are not only extensively shuffled between phage genomes, but have also been exchanged between phage and host genomes. For example, important components of the photosynthetic apparatus, such as *psbA* and *psbD*, have been discovered in the genomes of cyanophage *(16)*. Molecular phylogenetics played an important role in demonstrating that these genes are likely shuffled between *Synechococcus* and *Prochlorococcus* species via phage intermediates *(17, 18)*.

The purpose of this chapter is to introduce phage biologists to the field of molecular phylogenetics. The aim is to provide a general overview of the theory and methodology used to reconstruct phylogenetic trees from molecular sequence data. Most of the mathematical descriptions of methods have not been included herein. Instead, the reader is directed to the relevant literature on the subject. In addition, this chapter will not cover the construction of sequence alignments. However, the accurate assembly of sequence alignments is extremely important for proper phylogenetic inference and we suggest *(19, 20)* for further information. The chapter concludes with a brief discussion on how the evolutionary history of different genes from the same set of genomes can be compared, using T4-type phages as an example. In addition, a compilation of programs and packages that are available for conducting phylogenetic analyses is supplied as an accompanying appendix.

2 Phylogenetic Reconstruction

2.1 Terminology of Trees

A phylogenetic tree is a graphical representation of the evolutionary relationships among genes or organisms. In order to introduce some of the common terminology encountered in phylogenetics, a generic tree is shown in **Fig. 9.1**. Trees are composed of *nodes* and *branches*. A node represents a taxonomic unit in the tree and branches connect any two adjacent nodes. A *terminal node* (also called an operational taxonomic unit, OTU, leaf, or taxon) is located at the tip of the tree and, in molecular phylogenetics, represents an extant molecular sequence. An *internal node* (also referred to as a hypothetical taxonomic unit, HTU or ancestor) represents a hypothetical ancestor of the extant sequences. Any two terminal nodes, or any two groups of terminal nodes, that share a most recent common internal node can be referred to as *sister taxa* or *sister groups*. The overall branching pattern of a tree is called the *topology*. Generally, the *branch lengths* of a tree are proportional to the amount of evolutionary change that has taken place between each node on a tree. The amount of evolutionary change is proportional to the time since divergence and the rate of evolution. Two taxa can be linked by long branches because they diverged from one another in the distant past, or they may have diverged recently and the rate of evolution is high in one or both lineages.

A phylogenetic tree can be either rooted or unrooted. In the former case, the *root* is the common ancestor in the tree from which all other taxa in the tree have descended. A tree can be rooted with an *outgroup*, which should be a taxon known *a priori* to be phylogenetically outside the group of interest. The

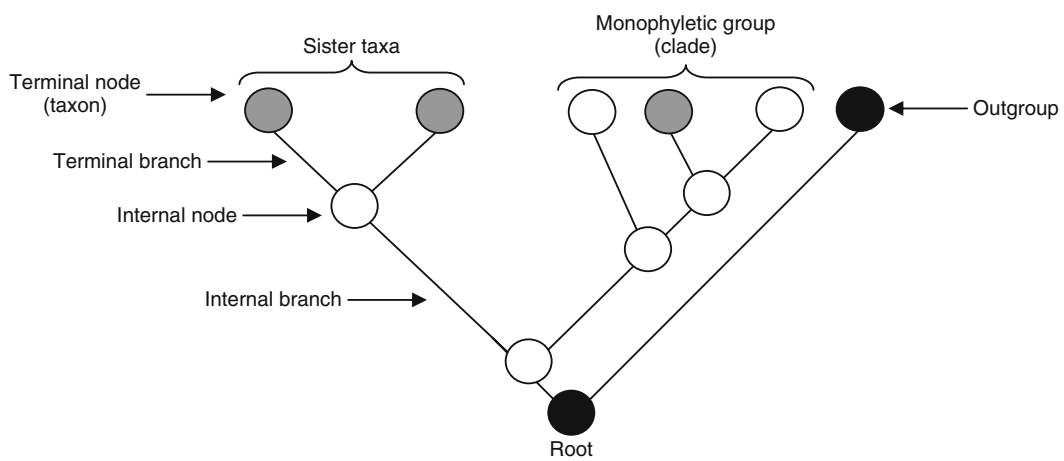


Fig. 9.1. Some common terminology used in describing a phylogenetic tree.

purpose of rooting a tree is to polarize the order of evolutionary events with respect to time. The direction from the root to the terminal nodes corresponds to evolutionary time. In terms of the number of bifurcations, the closer a node is to the root (deeper in the tree), the more ancient it is. Therefore, in a rooted tree, it is possible to define ancestor–descendent relationships. An unrooted tree, on the other hand, makes no claim as to the temporal direction of evolution or about ancestor–descendent relationships. In other words, the direction of time is unknown in an unrooted tree.

The placement of the root in a phylogenetic tree is usually based on some historical knowledge of the taxa under investigation. For example, the eukaryotic phylogeny is often rooted with an archaeal outgroup *(21)*. It is important to realize that the position of the root will influence the interpretation of evolutionary relationships. Take the unrooted five taxon tree pictured in **Fig. 9.2**, for example. Rooting the tree at position 1, produces a tree in which taxon C is more closely related to taxa D and E. However, if the tree is rooted at position 2, taxon C appears to be more closely related to taxa A and B. It is recommended that a tree be left unrooted if there is no specific reason for choosing any single root position. Therefore, no erroneous conclusions about the relationships between taxa will be drawn prematurely. However, it is often the case that a tree may be arbitrarily rooted on a group for display purposes only *(22, 23)*.

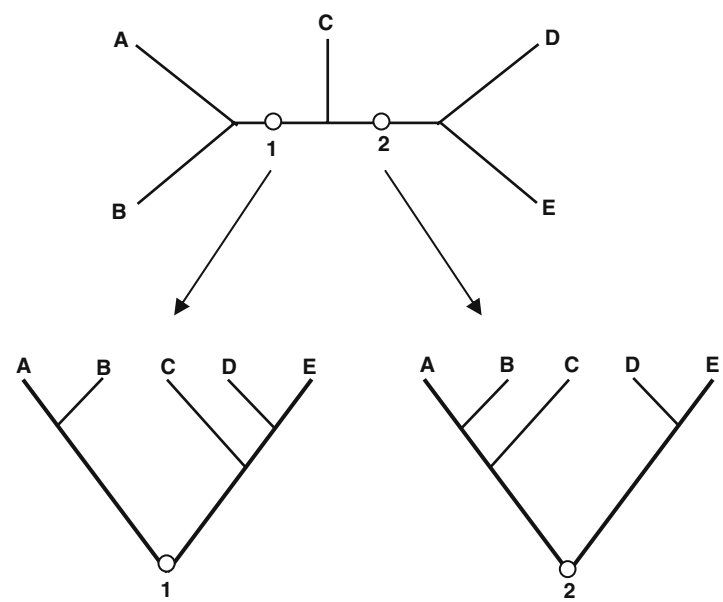


Fig. 9.2. Different positions of the root in a phylogenetic tree will lead to different conclusions about inferred evolutionary relationships. Taxon C appears to be more closely related to taxa D and E when the tree is rooted at position 1. If the tree is rooted at position 2, then taxon C emerges as a close relative of taxa A and B.

There are several important terms when referring to the evolutionary relationships between taxa in a rooted tree. A group of taxa is considered *monophyletic* if that group includes all the descendents of a single ancestral taxon (**Fig. 9.1**). A monophyletic group is also known as a *clade*. A non-monophyletic group, often referred to as *paraphyletic*, is one in which some of the descendents of the group have been omitted. The grey taxa in **Fig. 9.1** comprise a paraphyletic group. Paraphyletic relationships challenge our taxonomic classifications when members of an assumed group are phylogenetically unrelated.

Once a phylogenetic tree has been inferred, it is possible to differentiate between shared characters that are a result of descent from a common ancestor and those that are a result of independent evolution in two or more lineages. Some of the terminology surrounding this topic is presented in **Fig. 9.3**. A *synapomorphy* is a derived character state shared by two or more taxa because it was present in their common ancestor. *Homoplasy* occurs when a shared character state has evolved independently in two or more taxa. Homoplasy can result from several different evolutionary processes. *Convergent* or *parallel* evolution can produce the same character in two unrelated taxa. The difference between convergent and parallel evolution lies in whether the similar feature evolved from the same (parallelism) or different (convergence) ancestral condition. Homoplasy can also be caused by *secondary loss* of a derived character, leading to a reversion to the ancestral state. In addition to homoplasy and synapomorphy, there is *autapomorphy*, which is a derived character state unique to a single taxon (**Fig. 9.3**).

2.2 Terminology of Genes

Molecular phylogenetics involves the inference of evolutionary relationships through the comparison of molecular sequence data. A requirement is that the sequences under investigation be similar

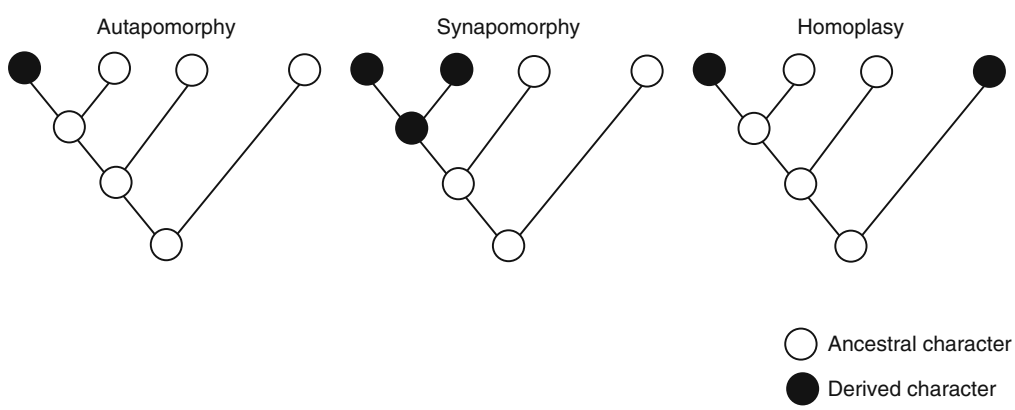


Fig. 9.3. Terminology used for describing characters on a phylogenetic tree. The terms above the three trees refers to the origin of the derived character state.

to one another because they descended from a common ancestor. Genes that have descended from a common ancestral sequence are referred to as *homologs*. There are several types of homologs and it is important to differentiate between them. *Paralogs* are homologous sequences that share a common ancestor due to one or more gene duplication events. *Orthologs*, on the other hand, are sequences that have diverged due to a speciation event. Differentiation between paralogs and orthologs is an important issue in phylogenetics when interpreting nodes in a tree. In theory, a phylogeny inferred from a set of orthologous sequences should, barring the effects of LGT, mirror the evolution of the organisms in which the orthologs are found. However, the unintentional use of several paralogs sequences can lead to a very different phylogenetic conclusion. For example, myoglobin and hemoglobin have a common ancestry and are the result of an ancient gene duplication. If one were to reconstruct a phylogeny for animals using a mixture of myoglobin and hemoglobin genes, then the tree topology would not accurately reflect the actual speciation events.

As stated earlier, the methodology behind collecting and aligning homologous sequences is not covered in this chapter. However, there is an important terminology with respect to molecular sequence data that will be used throughout the chapter. A *multigene alignment* consists of a series of homologous *sites* (or positions). Due to insertions and deletions of sites through evolutionary history, homologous sequences are not necessarily the same size and corresponding homologous sites in different taxa need to be found (i.e., need to be aligned). These sites are occupied by *characters*, which in the case of molecular data are either amino acids or nucleotides. The *character state* is defined by the identity of the character (i.e., the particular amino acid or nucleotide). In the amino acid alignment shown in **Fig. 9.4a**, we can say that at site 3 the character state for all five taxa is serine. Character states change by the process of *substitution* (also termed replacement). Again, in **Fig. 9.4a**, we can say that at site 7, there has been either one or several amino acid substitutions between serine and threonine.

2.3 Methods of Tree Reconstruction

Phylogenetic inference is a hypothesis-generating procedure, where an inferred tree represents the "best hypothesis" of evolutionary relationships based on the limited information contained in molecular sequence data and the assumptions of the phylogenetic reconstruction method (see below). Of the many possible evolutionary histories that could produce the observed differences between homologous sequences, we must have some method for choosing one or more best trees from all possible trees. There are two ways to approach this task in molecular phylogenetics. *Algorithmic methods* follow a fixed series of procedures (an algorithm) to derive a tree from the data. Algorithmic

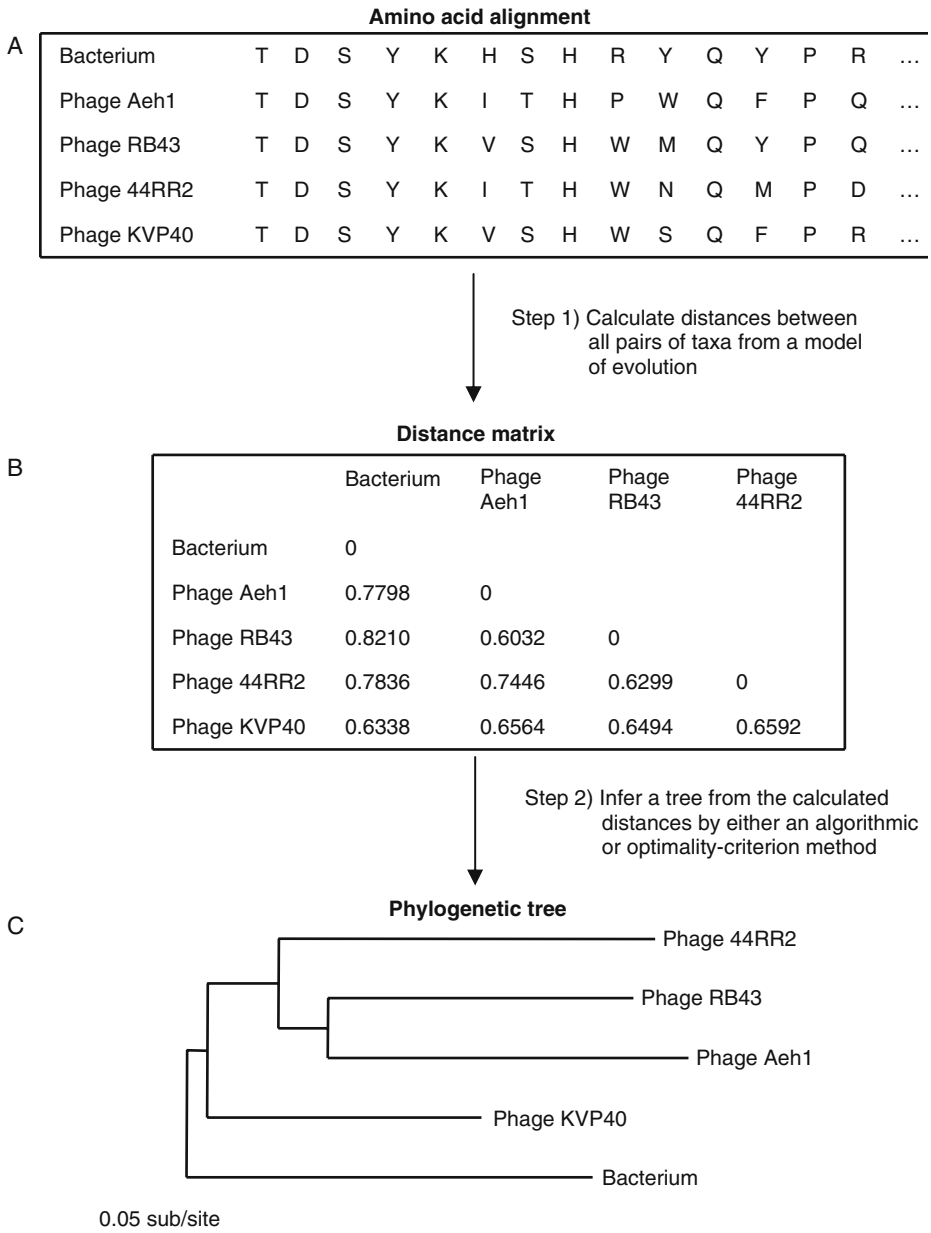


Fig. 9.4. An illustration of the two sequential steps involved in inferring a phylogenetic tree by the distance matrix method. From an edited amino acid alignment (**A**), a distance matrix is constructed by calculating the evolutionary distance between all pairs of taxa (**B**). Each value in the distance matrix corresponds to the distance between two taxa. For example, the distance between phage RB43 and phage Aeh1 is 0.6032 amino acid substitutions per site (sub/site). From the distance matrix, a phylogenetic tree can be inferred by a variety of methods (**C**). The tree was rooted on the bacterial sequence under the assumption that the divergence between the bacterial and phage sequences represents the most ancient node. The horizontal branch lengths are proportional to the amount of evolutionary change that has occurred between sequences. The length of the vertical branches has no evolutionary meaning. This particular phylogeny was inferred from an amino acid alignment of the *nadV* gene. Five homologous amino acid sequences were aligned with CLUSTALX *(87)*. The initial alignment was edited manually and ambiguously aligned sites were removed using MacClade 4.06 *(88)*. The distance matrix was calculated with PROTDIST using a JTT model of amino acid substitution. The tree was inferred from the distance matrix with NEIGHBOR using the neighbor–joining method. The PROTDIST and NEIGHBOR programs are part of the PHYLIP3.6 package *(89)*.

methods tend to be computationally fast. However, because they proceed directly to a final tree, without evaluating multiple trees, confidence in how well the algorithm-generated tree fits the data relative to an alternative hypothesized tree is unknown. Most distance-based clustering methods (e.g., UPGMA, neighbor-joining) fall into this category. *Optimality criterion* (Objective function) methods define a criterion for comparing alternative trees and then find the best tree that maximizes/minimizes the criterion. The advantage of optimality criterion methods, which include maximum likelihood and maximum parsimony methods, is they can define how good or bad any one tree is with respect to the criterion and the data. If many trees can explain the data equally well, the user will not be deceived into choosing a single tree as the best hypothesis. A tradeoff is that as the number of taxa becomes large (10 +), the number of possible trees becomes enormous (**Table 9.1**), and searching this "tree space" for the optimal tree can become computationally impossible. However, procedures exist for reducing the search time (e.g., heuristic search) and will be discussed below. Briefly, *Bayesian methods* are an alternative to the algorithmic and optimality criterion methods presented here *(24)*.

Table 9.1
Number of possible unrooted trees for increasing number of taxa[1]

Number of Taxa	Number of unrooted trees
3	1
4	3
5	15
6	105
7	945
8	10,395
9	135,135
10	2,027,025
11	34,459,425
12	654,729,075
13	13,749,310,575
14	316,234,143,225
15	7,905,853,580,625

[1] Data from Felsenstein 2004.

A critical point addressed throughout this chapter is that an inferred phylogeny is only as valid as the assumptions inherent to the phylogenetic reconstruction method and model of evolution. Hence, the implementation of an evolutionary model that accurately describes the true process of evolution is of crucial importance when estimating phylogenetic relationships. Below is an introduction to three commonly used phylogenetic methods: distance matrix, maximum parsimony, and maximum likelihood. Maximum likelihood (ML) is the most reliable at recovering phylogenetic relationships *(25, 26)*, but in the past its use was restricted due to long computational time. However, steadily increasing computer power has now brought ML to the forefront of molecular phylogenetics, and therefore methods and models using ML receive the most attention in this chapter.

2.3.1 Distance Matrix Methods

Phylogenetic inference by distance matrix methods involves two sequential steps (**Fig. 9.4**). First, the evolutionary distances (i.e., number of substitutions) between all taxa in an alignment is estimated based on a model of evolution. Then the results are tabulated in a distance matrix and one of a variety of approaches is used to reconstruct a phylogenetic tree from the pairwise distance values.

2.3.2 Estimating Pairwise Distances

One method to estimate evolutionary distances among taxa would be to simply count the number of observed pairwise substitutions between taxa in the alignment, since it is expected that the more distantly related any two taxa are, the more changes will have occurred between their sequences. This is indeed true, however the estimated distances might be incorrect because not all substitutions are observable. Multiple substitutions at the same site in the same taxon go uncounted, leading to an underestimation of the true evolutionary distance between taxa (**Fig. 9.5**).

Recovering an accurate distance between two sequences requires a model of evolution that will correct for these unobservable changes. The simplest such model for protein sequence evolution is the *Poisson model*, which is analogous to the Jukes–Cantor (JC) one-parameter model of DNA sequence evolution *(27)*. The Poisson model assumes substitutions between each of the 20 amino acid residues occur with equal probability and corrects for multiple substitutions by incorporating the probability that unobserved substitutions of sequential amino acid replacements (e.g., ser->thr->ser) have occurred at the same site.

The Poisson model is an oversimplification of the evolutionary process. In addition to equal substitution probability, it makes several other assumptions including (1) an equal frequency of all amino acids, (2) an equal evolutionary rate at all sites, and (3) independent evolution between sites. Empirical observations of protein alignments demonstrate these assumptions are often

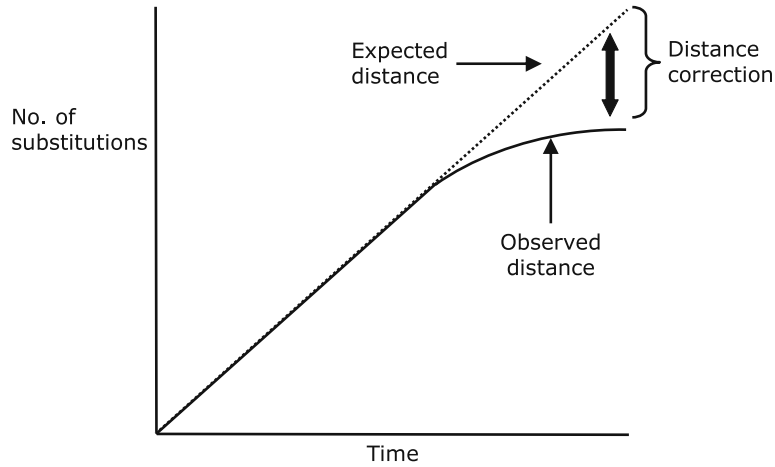


Fig. 9.5. The relationship between time and the observed distance among two sequences (*solid line*) is not linear, but instead forms a plateau. The plateau is due to the accumulation of multiple substitutions at the same site, which are unobservable. As such, the observed distance is an underestimation of the true distance between sequences. Recovering an accurate estimation of the expected distance (*dotted line*) between two sequences requires a model of evolution that will correct for the unobservable substitutions.

not met in nature. For example, amino acid substitution tends to occur much more frequently between amino acids of similar physiochemical properties *(19)*. Therefore, much more realistic models of protein evolution have been devised (see section on Models of Evolution).

2.3.3 Inferring a Tree from a Distance Matrix

There is an extensive variety of both algorithmic and optimality criterion methods available for inferring a phylogenetic tree from a matrix of evolutionary distances. The simplest algorithmic method is Unweighted pair-group method with arithmetic mean (*UPGMA*) *(28)*. UPGMA uses a sequential clustering algorithm to group taxa in order of decreasing similarity. UPGMA makes the assumption that there is a linear relationship between evolutionary distance and divergence time, or, in other words, that the rate of evolution is equal and has remained constant among taxa (i.e., ultrametric or clock-like). This assumption is rarely, if ever, met and therefore it is advised that UPGMA not be used to infer a best tree. There are many other superior methods for tree reconstruction that are as easy to implement and are computationally fast.

One such widely employed algorithmic procedure that does not make the assumption that data are ultrametric is the *neighbor-joining method (29)*. Neighbor-joining is a star decomposition algorithm that attempts to minimize the overall branch length of the tree. From an initial star tree with a single internal node, all possible two node trees are constructed, where the second

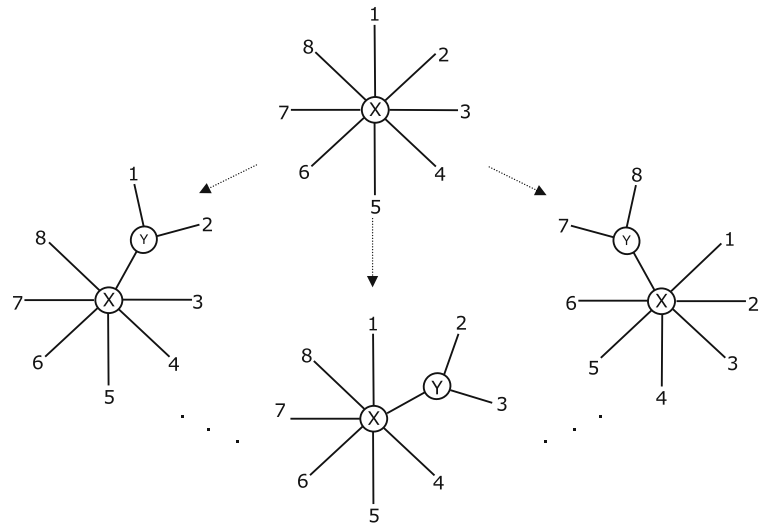


Fig. 9.6. The initial steps in reconstructing a tree by the neighbor-joining (NJ) method. NJ is a star decomposition method that begins with a star tree, which is a tree where all taxa (8 in this example) are connected by a single node (X). The NJ algorithm constructs all possible two node trees. The second node (Y) represents all possible taxa pairings. The pair of taxa that results in the tree with the smallest sum of branch lengths is chosen as the first pair of "neighbors." These two taxa are treated as a single composite taxon, the distance between the composite taxon and all other taxa is computed and the new distances are used in a second round of locating neighbors by identifying the tree with a minimum overall branch length. This process is repeated sequentially until a fully resolved tree is assembled. Modified from *(90)*.

node consists of all pairs of taxa (**Fig. 9.6**). The pair of taxa that gives the tree with the smallest sum of branch lengths (S) is chosen as the first pair of "neighbors." These two taxa are then treated as a single composite taxon, a new distance matrix is computed and the process is repeated successively until a fully resolved tree is assembled *(30)*. Modified versions of the original neighbor-joining method, such as BioNJ *(31)* and Weighbor *(32)*, have been formulated and they tend to outperform the original neighbor-joining algorithm *(33)*. Because of fast run times, neighbor-joining is particularly useful for large studies or bootstrap resampling studies that require analysis of multiple datasets (see section on Nonparametric Bootstrap Analysis).

Optimality-criterion methods attempt to fit a tree to a distance matrix, usually by minimizing some criterion. *Minimum evolution* is a straightforward example of such a method. As the name suggests, the objective of minimum evolution is to minimize the sum (S) of all branch lengths in a tree *(34, 35)*. In this manner, neighbor-joining is an algorithmic approximation of minimum evolution. However, unlike minimum evolution, neighbor-joining constructs only one tree instead of searching through "tree space" for an optimal tree. A second type

of method, the *least squares method* attempts to minimize the difference between the observed distances in the matrix and the expected distances given in the hypothesized tree *(35)*. A widespread least squares method currently in use is the Fitch–Margoliash method *(35, 36)*.

2.3.4 Maximum Parsimony

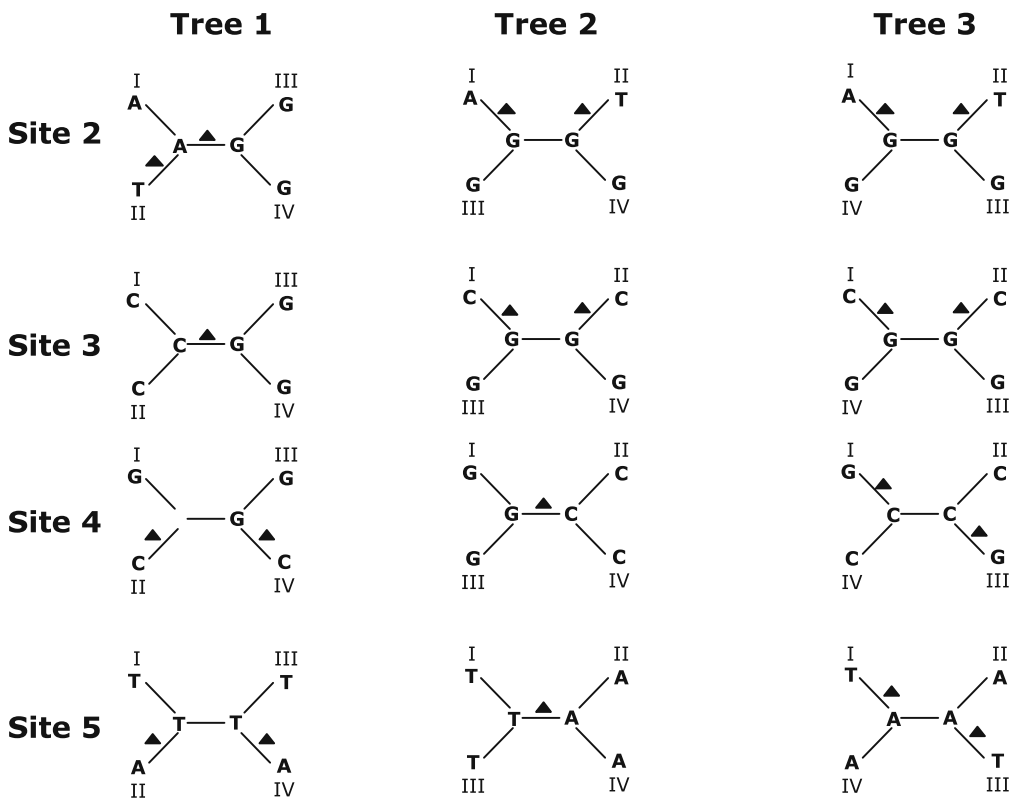
Parsimony methods were among the first for inferring phylogenies and are based on the concept that the best hypothesis is the one that requires the least amount of evolutionary changes *(37)*. In molecular phylogenetics, the principle of parsimony is to find the tree (i.e., hypothesis) that requires the minimum number of substitutions to explain the observed/inferred difference between taxa. Maximum parsimony (MP) is thus an optimality-criterion method in which the criterion (i.e., number of substitutions) is to be minimized. The tree that minimizes the number of substitutions required to explain the data is called the *maximum parsimony tree*.

Parsimony begins with the classification of sites as either informative *sensu* parsimony or uninformative. A site is considered informative if it favors a subset of trees over all possible trees. The classification of sites and the basic procedure behind choosing the maximum parsimony tree can be illustrated by considering the hypothetical four taxa alignment in **Fig. 9.7a**. For four taxa, there exists three possible unrooted trees and we can use the information in the sequence alignment to choose which tree amongst these three possibilities is the most parsimonious. Site 1 is considered uninformative because all sequences in this site have the same character state (adenosine) and no change is required, regardless of the inferred tree. Site 2, although not invariant, is also uninformative because the most parsimonious explanation for this character pattern requires an identical minimum of two substitutions in all three trees (**Fig. 9.7b**). The remaining three sites are all parsimony informative, in that they favor one tree over the other two possibilities, and can be used to search for the maximum parsimony tree. To identify the maximum parsimony tree, we first calculate the minimum number of substitutions at each informative site for all three possible trees (**Fig. 9.7b**). For example, a single change is required to explain the substitution pattern observed at Site 3, given Tree 1, whereas a minimum of two substitutions must be inferred from Trees 2 and 3 in order to reconstruct the same substitution pattern. Therefore, Tree 1 is the most parsimonious explanation for the observed substitution pattern at Site 3. Next, we sum the number of substitutions across all informative sites for each possible tree and select the tree that gives the minimum number of changes (**Fig. 9.7c**). In our example, Tree 2 is the most parsimonious tree because it infers only six substitutions, whereas Trees 1 and 3 infer 7 and 8 substitutions, respectively. This is a simple example of parsimony, presented here

A

Sequence	Site				
	1	2	3	4	5
I	A	A	C	G	T
II	A	T	C	C	A
III	A	G	G	G	T
IV	A	G	G	C	A

B



C

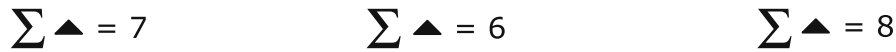


Fig. 9.7. Phylogenetic inference using the maximum parsimony (MP) method from (**A**) a hypothetical nucleotide alignment consisting of four taxa. (**B**) There are three possible unrooted trees for four taxa and the nucleotide alignment can be used to find which of the three possible trees is the maximum parsimony tree. For each variant site (sites 2–5), the minimum number of substitutions required to explain the character pattern is calculated for each of the three possible trees. Terminal nodes are labeled with the taxon identity (I–IV) and the nucleotide character at the site under consideration. Internal nodes are labeled with a most parsimonious ancestral character state. Nucleotide substitutions along the branches are indicated by a ▲ symbol. (**C**) The maximum parsimony tree in this hypothetical example is Tree 2 because it minimizes the number of inferred substitutions. Modified from *(90)*.

to explain the basic principle. However, different types of parsimony exist and for further information, see *(37)*.

2.3.5 Inconsistency in Parsimony

Parsimony is a method that endeavors to minimize the number of substitutions inferred by the tree, and in doing so also minimize the number of homoplasies. When the amount of divergence between taxa is small, we predict that little evolutionary change has occurred, and the amount of homoplasy to be low. However, when divergence between taxa is large, homoplasies can become much more common in the data and this can present a significant problem for phylogenetic reconstruction by maximum parsimony. This is particularly problematic for taxa with an increased rate of evolution.

Suppose the true phylogeny for a collection of four taxa is as shown in **Fig. 9.8a**, where the length of the branches specify the amount of evolutionary change that has occurred. The long terminal branches suggest that the rate of evolution is accelerated in taxa I and II. The probability that homoplasious substitutions have occurred in these fast-evolving lineages is higher compared to taxa III and IV. Hence, an informative site may have a substitution pattern as shown in **Fig. 9.7b**. Unfortunately, this pattern supports an incorrect tree if incorporated into parsimony (**Fig. 9.7c**). Because this inconsistency in parsimony tends to cluster long branches together, it has become known as "long branch attraction" *(38)*. Long branch attraction can be a problem in all phylogenetic methods and will be discussed more thoroughly in a later section.

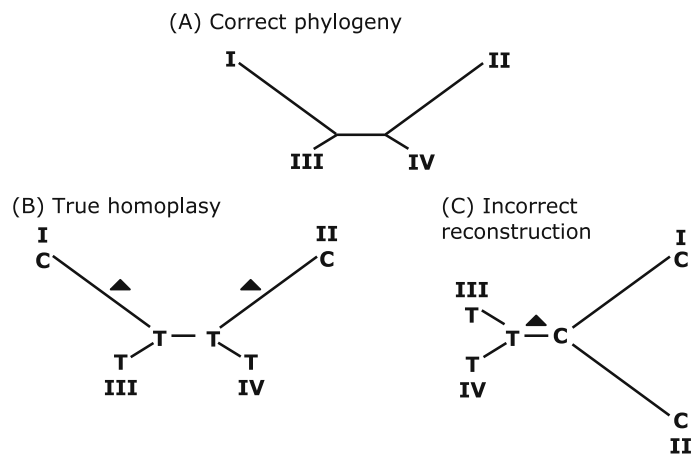


Fig. 9.8. An illustration of the "long branch attraction" artifact in maximum parsimony. (**A**) The true phylogeny for four hypothetical taxa. The rate of evolution is elevated in taxa I and II relative to taxa III and IV. (**B**) Parallel evolution in the fast evolving taxa can result in two independent substitutions (▲) of T to C in taxa I and II. (**C**) The most parsimonious reconstruction infers a single substitution along the internal branch, which leads to the incorrect reconstruction of the two longest branches as neighbors.

2.3.6 Maximum Likelihood

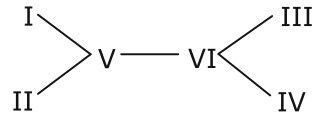
Maximum likelihood is an optimality-based method, which evaluates a hypothesized tree in terms of the probability that it would lead to the observed sequence data under a proposed model of evolution *(39,40)*. The principle of maximum likelihood is to find the tree that maximizes the likelihood value for the data.

Calculating the likelihood of a tree is a statistical procedure, which, like parsimony, considers each site of the alignment individually. Here we will introduce the basic principles of the likelihood calculation using a collection of aligned nucleotide sequences and a four taxon tree (**Fig. 9.9**). If we wanted to calculate the likelihood for the unrooted tree shown in **Fig. 9.9b**, we begin by evaluating each site individually, and then combine all the site likelihoods together into a total likelihood value for

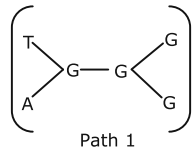
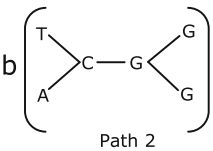
A

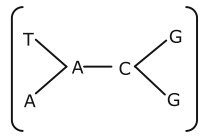
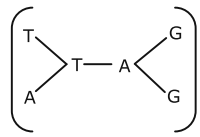
Sequence	Site					
	1	2	3	4	…	N
I	A	T	C	G	…	T
II	A	A	C	C	…	A
III	A	G	G	G	…	T
IV	A	G	G	C	…	A

B



C

$L_{(\text{Site 2})} = \text{Prob}$  Path 1 $+ \text{Prob}$  Path 2 $+$

Prob  Path 3 $+ \cdots +$  Path 16

D

$$L = L_{(1)} \bullet L_{(2)} \bullet \ldots \bullet L_{(N)} = \prod_{J=1}^{N} L_{(j)}$$

E

$$\ln L = \ln L_{(1)} + \ln L_{(2)} + \ldots + \ln L_{(N)} = \sum_{J=1}^{N} \ln L_{(j)}$$

Fig. 9.9. Phylogenetic inference using the maximum likelihood method. See the text for a description of the procedure used to assess the likelihood of the data for a hypothesized tree. Modified from *(27)*.

the tree. To calculate a site likelihood, for example, the likelihood of Site 2 in the nucleotide alignment shown in **Fig. 9.9a**, we must consider every potential path of evolution that could have led to the extant character pattern. There are a total of 16 possible paths to consider for Site 2 (**Fig. 9.9c**) and, although some pathways are less reasonable than others, they must all be considered because they exist with a probability greater than zero and contribute positively to the likelihood value of the tree. The likelihood for each site is found by summing the probabilities of each of the 16 possible pathways (**Fig. 9.9c**). Under the assumption of independent evolution of sites, the overall likelihood for the tree is equal to the product of the likelihoods for each site (**Fig. 9.9d**).

In practice, likelihood values are extremely small. Therefore, it is the logarithmic transformation of the likelihood that is usually evaluated. When the log likelihood (lnL) is considered, multiplication is transformed to summation and therefore the equation in **Fig. 9.9d** is transformed to that in **Fig. 9.9e**.

2.3.7 An Example of Maximum Likelihood

To demonstrate the use of maximum likelihood in phylogenetic reconstruction, we investigated the evolution of an interesting gene recently discovered in the genome of vibriophage KVP40. The gene is *nadV* and encodes a component of the pyridine nucleotide (NAD^+) salvage pathway *(41)*. This gene is common among cellular organisms but its presence in a phage genome was a novel finding at the time the KVP40 genome was analyzed. Our purpose was to determine whether other phage genomes possess a *nadV* gene and to investigate the origin of this "cellular gene" in KVP40. We identified 263 sites along the NadV amino acid alignment that were conserved enough for phylogenetic analysis. The maximum likelihood tree inferred from these 263 sites is presented in **Fig. 9.10**; all programs used for the analysis are described in the figure legend. A discussion on the model of evolution used is presented in the following section.

From this tree, it is clear that there are at least seven phages that contain a *nadV* homolog in their genomes and which infect a diverse host range including both Gram-negative and Gram-positive bacteria. In addition, it is apparent from this phylogeny that KVP40 shares a recent common ancestor for this gene with three other phages. In other words, these four phages are monophyletic for the *nadV* gene. In contrast, the remaining three phages are paraphyletic in that they are distributed amongst bacterial NadV sequences. These results suggest the seven NadV phage sequences do not share a common phage ancestor, but instead were independently exchanged with bacteria multiple times.

2.4 Models of Evolution

The importance of an accurate evolutionary model for reconstructing phylogenetic relationships from molecular sequence data cannot be understated. In the above section, we outlined

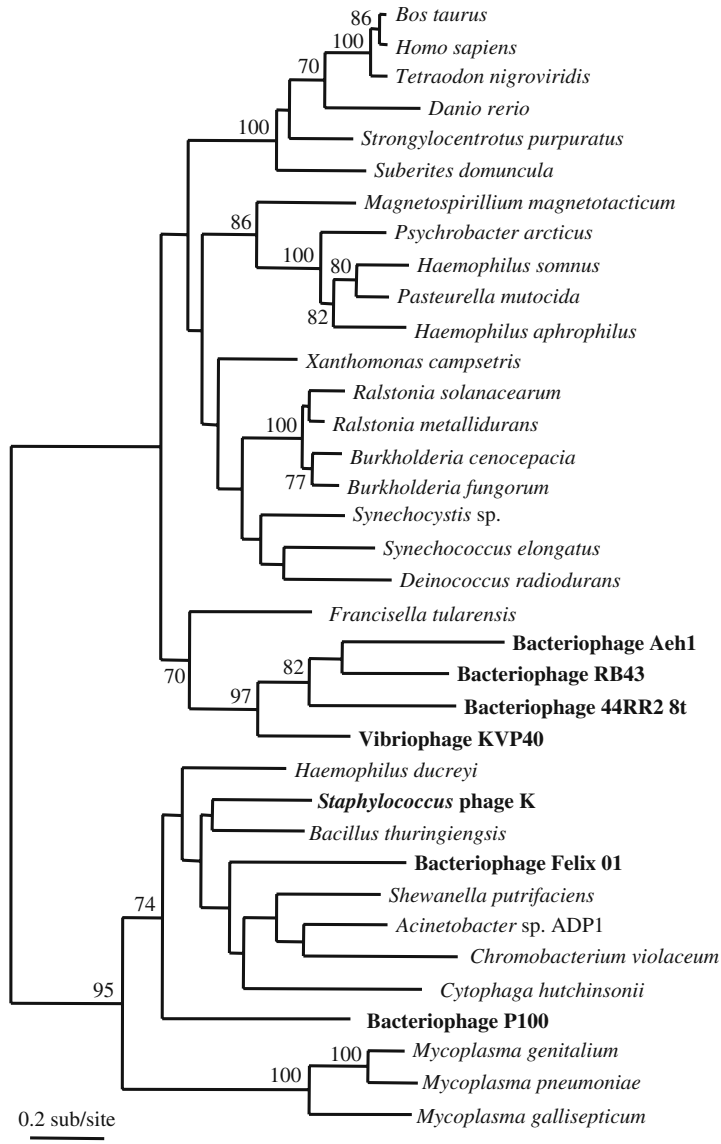


Fig. 9.10. Best ML tree inferred from the NadV amino acid alignment using PHYML *(91)*. The evolutionary model and parameters employed in the analysis were a WAG substitution matrix+Γ (eight categories)+I, α =1.598, I=0.087. Support for nodes corresponds to bootstrap values for 100 pseudo-replicates. The 100 pseudo-alignments were constructed with SEQBOOT. Pseudo-trees were inferred with same method as the best tree and a consensus tree was constructed by the majority rule method using CONSENSE. SEQBOOT and CONSENSE are part of the PHYLIP 3.6 package *(89)*. Only bootstrap values greater than 70% are displayed on the tree. Bacteriophage NadV sequences are highlighted in bold. The tree has been arbitrarily rooted for display purposes only.

several common methods for inferring phylogeny. However, we have yet to describe the models that are usually employed when calculating evolutionary distances between taxa or evaluating the likelihood of a hypothesized tree. The function of a model of evolution is to reproduce the evolutionary process. The parameters included in a model are chosen to reflect some aspect of the evolutionary process. Model parameters include a substitution matrix, which incorporates the different probabilities of change between character states. Models also incorporate character frequency parameters (i.e., how often each nucleotide or amino acid is observed in the data) and relative evolutionary rate parameters, which, as we will see below, estimate the true variation in evolutionary rates at different sites along an aligned molecular sequence. Model parameters can either be fully defined *a priori*, or be estimated from the dataset under study *(27)*.

The tradeoff for increasing model complexity is usually increased computational time. Hence, preliminary phylogenetic analyses are often performed using simple models of evolution, so as to generate a rough estimate of relationships that can be further tested with more complex models. Before a conclusion is drawn from a phylogenetic analysis, it is important to investigate how the results vary with both the phylogenetic method and the model *(42, 43)*. In some cases, violation of the assumptions inherent to a model of evolution can have a devastating effect on the results (see section entitled Systematic error and long branch attraction).

2.4.1 Substitution Matrices

Nucleotide substitution models are theoretically derived, and can range from the very simple (JC; equal rate of substitution for all nucleotides) to the very complex (General time reversible; a distinct rate for each substitution type) *(27)*. Statistical methods, such as the likelihood ratio test of model fit, exist that search for the nucleotide substitution model that best fits the data without introducing unnecessary model complexity *(44)*.

With respect to amino acids, earlier we described the Poisson model as a simple, theoretical model of amino acid substitution that treats all amino acid replacements with an equal probability. However, empirical evidence has revealed that amino acids are much more likely to be replaced by amino acids with similar physiochemical properties (polarity, size, and charge), then is assumed under an equal replacement probability model, such as the Poisson model *(19)*. This observation has led to the implementation of empirically based models of amino acid substitution, which consist of a 20×20 rate matrix that estimates the probabilities for each amino acid being replaced by each alternative amino acid. This approach was first employed by Dayhoff and co-workers who developed an amino acid substitution matrix by calculating the replacement probabilities of amino acids from trees inferred

from protein alignments by maximum parsimony *(45)*. Additional empirical models of amino acid substitution have been developed. The Jones–Taylor–Thornton model (JTT) is based on a more up to date substitution matrix constructed from a larger database of sequences *(46)* and as such is preferred over the Dayhoff model. In the JTT model, the probabilities were estimated from protein alignments that were at least 85% identical, in order to reduce the chance that an observed change resulted from multiple substitutions *(46)*. The PMB model is derived from the BLOCKS database of conserved protein motifs *(47)*. The NadV maximum likelihood phylogeny presented in **Fig. 9.10** was inferred from the WAG substitution model, which uses a substitution matrix calculated using an approximate maximum likelihood method *(48)*.

2.4.2 Rate heterogeneity

All models described above assume that every amino acid site evolves at the same rate. This is not a valid assumption as substitution rates can vary extensively along a protein due to differences in functional constraint across amino acid sites. Maximum likelihood analysis becomes inconsistent under the assumption of equal rates when the evolutionary process exhibits variation in evolutionary rate from site to site, also known as *among-site rate variation* (ASRV).

The most commonly used distribution for modeling ASRV is the *gamma* (Γ) *distribution (49)*. The shape of the Γ distribution is defined by the shape parameter, alpha (α) (**Fig. 9.11a**). The α value is inversely related to the variation in the substitution rate and will be low if variation in the substitution rate among sites is large. As alpha increases, the variation in the substitution rate among sites decreases until, as alpha approaches infinity, it converges on an equal-rates model.

Incorporating a continuous distribution, such as the Γ distribution, into a likelihood model can be computationally intensive *(49)*. To reduce computation time, an alternative method was devised, where the Γ distribution is divided into several discrete rate categories of equal probability, and the median of each category is used to represent all of the rates in the category. Amino acid sites are then assigned to one of the relative rate categories. Typically, this discrete approximation of the Γ distribution is estimated by either four or eight relative rate categories *(50)*. An example of approximating a Γ distribution with four rate categories is shown in **Fig. 9.11b**, where the area under the curve (probability) has been divided into four equal parts, each part consisting of one of the four rate categories.

Mixed models of rate heterogeneity are also often employed in phylogenetic analyses. For example, a Γ distribution (eight rate categories) + Invariant model was implemented in the inference of the NadV phylogeny (**Fig. 9.10**). This means that in addition to the Γdistribution, the model has an extra level of complexity

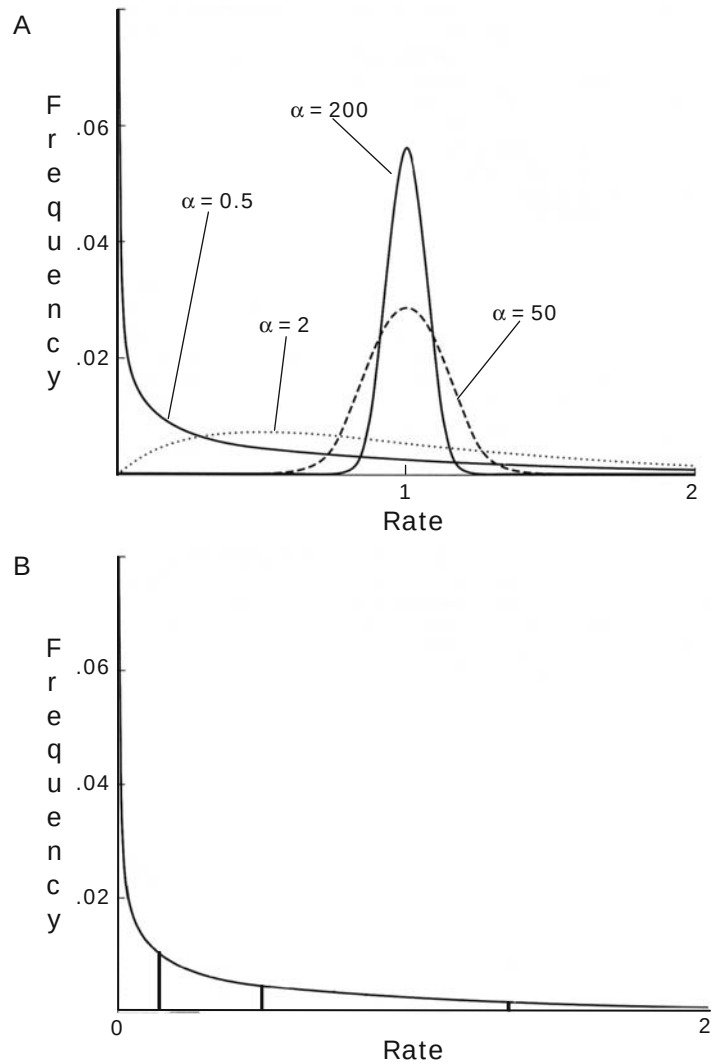


Fig. 9.11. (**A**) The gamma (Γ) distribution for different values of the shape parameter (α). The variation in the substitution rate among sites is inversely related to α. When α is small, the variation in substitution rate is high. As α increases, the variation decreases until, at $\alpha = \infty$, it converges on an equal-rate model. (B) The continuous Γ distribution can be approximated by division into several discrete rate categories of equal probability. Here a Γ distribution with $\alpha = 0.5$ has been approximated by four rate categories. The vertical lines under the curve correspond to the boundaries between the four categories. Modified from *(27)*.

because it has an additional parameter that solely accounts for the proportion of invariant sites in the alignment. This is the model we recommend for incorporating rate heterogeneity.

2.5 Searching Tree Space

Optimality-criterion phylogenetic methods, such as maximum likelihood, search through multiple trees to find the optimal tree. The most thorough approach is to evaluate all possible trees in an

exhaustive search, and choose the globally optimal tree. Exhaustive search methods are only useful for a small number of taxa since the number of possible trees increases rapidly with increasing number of taxa, such that it becomes computationally impossible to evaluate all trees (**Table 9.1**).

One approach to overcoming this problem is to employ *heuristic search* methods. Heuristic tree searches seek the optimal tree though the use of iterative trial and error processes, which examine a subset of all possible trees. Most searches of this type operate under similar principles, where a starting tree is first constructed by a fast algorithmic tree building method, such as neighbor joining. Alternative trees are then examined by systematically rearranging branches and those that have a higher optimality than the reference tree are maintained and used as the new reference tree. This process is repeated until there is failure to uncover a better tree. There are many branch swapping algorithms in use for generating alternative topologies. Several examples are briefly outlined here. *Nearest neighbor interchange* (NNI) represents a branch swapping method that results in local rearrangements of a tree topology *(27)*. In *subtree pruning and regrafting* (SPR), all possible subtrees are "pruned" from the reference tree and then "regrafted" at an alternative location *(27)*. *Tree bisection and reconnection* (TBR) considers all possible bisections of a tree, from which all combinations of pairwise reconnections are evaluated *(27)*. The tradeoff of heuristic searching is that this method is not guaranteed to find the globally optimal tree due to the possibility of local optima in the evaluated criterion (see *(37)*).

2.6 Error Associated with Inferred Trees

There are several sources of error encountered when inferring a phylogenetic tree. *Random error* is the deviation from the true tree, because there is a limited length of sequence data. Random error will therefore tend to decrease with an increasing length of data, as the stochastic variation associated with small sample size becomes less. *Systematic error* is the deviation from the true tree due to incorrect assumptions in the method or model used for phylogenetic inference. Systematic error will introduce a bias that may support the wrong tree and, unlike random error, the addition of more data will tend to increase support for the incorrect tree. Fortunately, procedures for assessing both random and systematic error exist and considerable effort has been directed at minimizing their effects in phylogenetic inference. Several of these methods will be described here, yet, for further discussion on phylogenetic artifact due to phylogenetic pitfalls see *(51)*.

2.6.1 Random Error

When the length of sequence data available for a given set of taxa is limited, there exists the possibility that one tree will be favored over a second tree by chance alone. This random error associated

with finite sample size will only disappear once an infinite amount of data has been obtained, a realistically unattainable situation. Hence, once the best tree has been generated, it is important to assess how sensitive this tree is to the amount of sequence data from which it was inferred.

2.6.2 Non-parametric Bootstrap Analysis

The non-parametric bootstrap analysis is a statistical technique that uses random resampling of data with replacement to determine sampling error or the confidence interval for an estimated parameter, in this case the groups of the hypothesized best tree *(52)*. This type of analysis is commonly employed in phylogenetics and is outlined in **Fig. 9.12**. To begin, a series of pseudo-alignments of the same length as the original alignment are generated by a sample with replacement procedure from the original alignment. Sites can be sampled multiple times with the same probability, or not at all. Typically, either 100 or 1,000 pseudo-alignments are generated, depending on the computational time required by the phylogenetic method. From each pseudo-alignment, a tree is inferred, resulting in a collection of 100 or 1,000 estimated trees. The phylogenetic information (i.e., the number of trees in which the same group of taxa is recovered) contained in this set of trees is summarized in a consensus tree. There are several different methods for constructing a consensus tree, but the most common are strict consensus and majority-rule consensus trees *(37)*. In practice, bootstrap values > 70% are

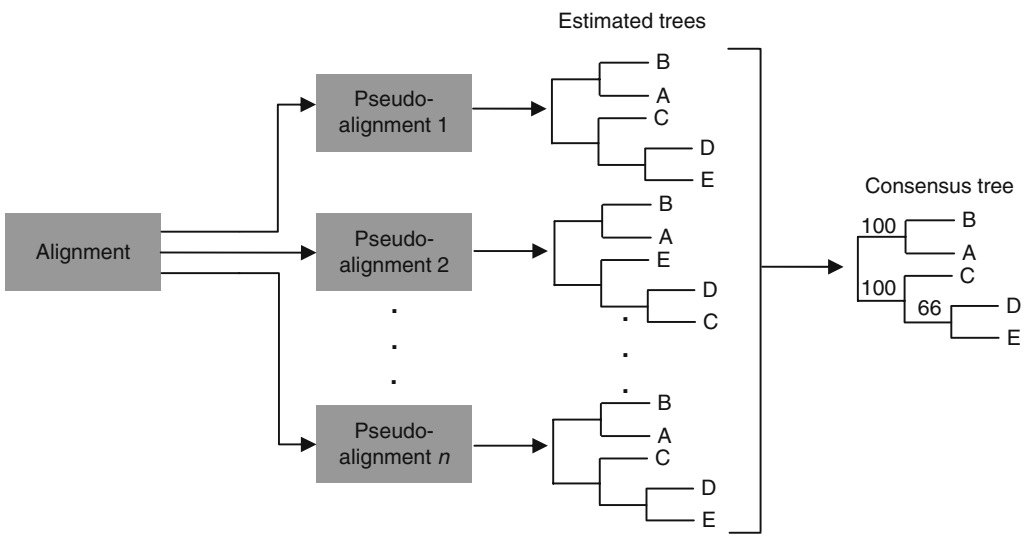


Fig. 9.12. The non-parametric bootstrap analysis. From the original alignment, *n* pseudo-alignments of the same length are generated by a sample with replacement procedure. From each pseudo-alignment a tree is inferred, resulting in *n* trees. Typically, n = 100 or 1,000 depending on the computational time required by the phylogenetic method. The percentage of times phylogenetic groups are recovered in the *n* trees is summarized by calculating a consensus tree. The values on the branches correspond to the percentage of trees in which that branch was reconstructed.

often considered significant support for a clade, however, the significance of bootstrap values is highly debated. In our example, the clade containing A and B is reconstructed in all estimated trees, therefore this branch is given a value of 100%. In contrast, the clade containing D and E is only recovered in 66% of the estimated trees and therefore our confidence that this is a robust relationship is lower. The recovery of this clade may reflect a weak phylogenetic signal, potentially affected by stochastic error in the data.

2.6.3 Evaluating Alternative Topologies

In maximum likelihood, the optimal tree may not be significantly better than a multitude of other hypothesized trees. Several methods have been developed for assessing the confidence that the optimal tree is significantly better than alternative trees. These include the Kishino–Hasegawa (KH; *(53)*), the Shimodaira–Hasegawa (SH; *(54, 55)*), and the Approximately Unbiased tests (AU; *(56)*). Although statistically different, these methods all follow the general procedure outlined in **Fig. 9.13**. The investigator provides a series of alternative topologies to test if they are significantly less likely to explain the observed data than the optimal tree. For each tree, a probability value (*P*-value) is calculated, which represents the confidence that this tree is the true tree. Trees are often rejected at $P < 0.05$. Due to method-specific bias, the KH test often rejects many plausible trees, which can lead to preference of the wrong tree *(54,55)*, while the SH test risks being

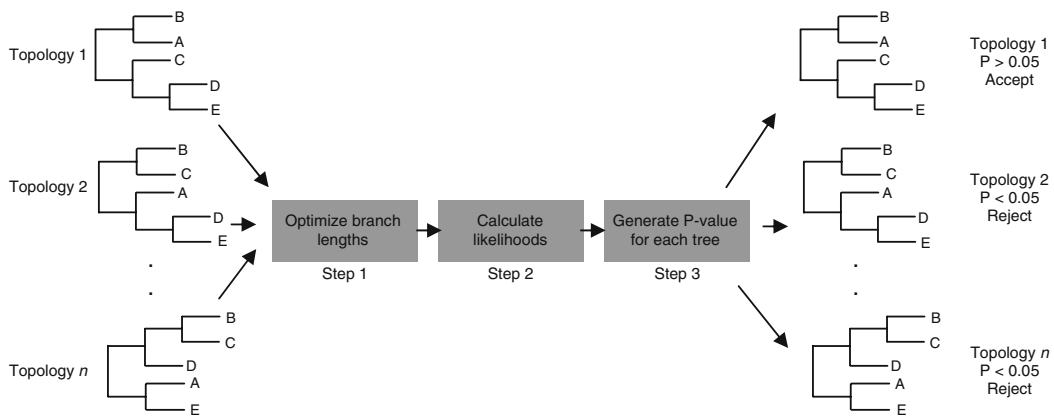


Fig. 9.13. An illustration of the approach used to evaluate the confidence in a phylogenetic tree. A series of hypothesized topologies is supplied by the user. These topologies are entered into a phylogenetic program that will optimize the branch lengths of the topologies and estimate the likelihood of each site of the given gene and the global likelihood of the data for each tree (Steps 1 and 2). Programs that can perform such an analysis are PUZZLE *(92)*, PAUP *(93)*, PAML *(94)*, and MOLPHY *(95)*. The results are then used as input for a program, such as CONSEL *(96)* that can evaluate the trees with respect to the likelihood values that were generated. CONSEL will provide *P*-values for the AU, KH, and SH tests. When the *P*-value < 0.05, the tree is considered significantly worse at explaining the observed molecular data than the other trees and is rejected as a plausible topology. In our hypothetical example, all the topologies except for Topology 1 have a *P*-value < 0.05, indicating that Topology 1 is the best tree and that the alternatives are all significantly worse.

too conservative by failing to reject many plausible trees *(56)*. The AU test is the latest test of tree selection and is considered as the least biased method *(57)*.

To demonstrate how to test which parts of a tree are robust, we can use the AU test to investigate the significance of certain relationships in the NadV maximum likelihood tree previously discussed (**Fig. 9.10**). In the NadV phylogeny, there are four phages (KVP40, Aeh1, RB43, and 44R2 8t) that form a well-supported (97% bootstrap) monophyletic group, indicating they share a common ancestor to the exclusion of cellular organisms. Contrastingly, the three remaining phages (K, Felix 01, and P100) are paraphyletic and do not appear to share an exclusive common ancestor.

Perhaps there were some biological reasons to believe these three phages (K, Felix 01, and P100) should form a monophyletic group that branches after the Mycoplasmatales. The AU test can be applied to evaluate the hypothesis that rearranged trees in which the bacteriophages are monophyletic and branch after the Mycoplasmatales are not significantly worse than the original tree. For this demonstration, we have chosen three rearranged test topologies (**Fig. 9.14**). The only difference between the three trees is the branching order of the bacteriophage sequences. *P*-values are generated for the original and each rearranged tree using the method summarized in **Fig. 9.13**. As expected, the original tree has a *P*-value near to one ($P = 0.995$), while the rearranged trees are all rejected ($P < 0.05$). Therefore, the rearranged trees are significantly worse at explaining the data than the original tree, and the NadV phylogeny may be at odds with the supposed biological data that lead us to originally propose these bacteriophages should be monophyletic.

2.6.4 Systematic Error and Long Branch Attraction

Systematic error occurs when the assumptions of a phylogenetic method are violated by the data. Such is the case when the real evolutionary process is not accurately described by oversimplified models of evolution. The long branch attraction (LBA) artifact is the most widely discussed symptom of systematic error in phylogenetic reconstruction. It arises because our models of evolution fail to capture the genuine variability in evolutionary rate among taxa *(51)*. LBA tends to cluster the longest branches together in a tree, irrespective of the true relationship between taxa under investigation *(38)*. When a distant outgroup (itself a long branch) is present, LBA can result in the attraction of fast evolving taxa to the base of the ingroup. The most notorious example of LBA is the microsporidia, which were first thought to represent an ancient eukaryotic lineage because of their early emergence in eukaryotic phylogenies *(58, 59)*. The microsporidia were later deemed to be highly diverged fungi *(60, 61)*, which were artifactually attracted to the long-branched archaeal outgroup because of improper modeling of evolutionary rate variation *(62, 63)*.

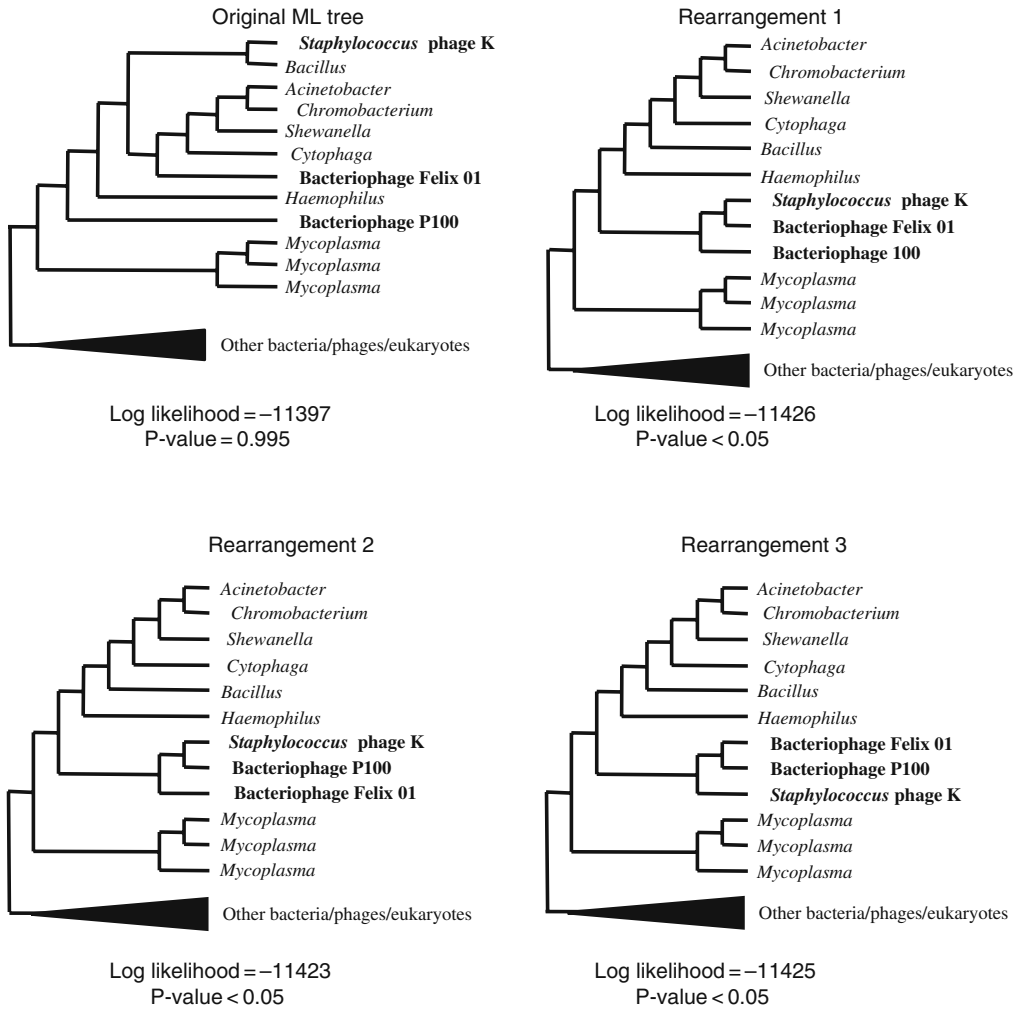


Fig. 9.14. An example of using the AU test to evaluate topological rearrangements of the NadV ML tree (see **Fig. 9.10**). Three rearrangements of the NadV ML tree were compared to the original ML tree. In all three rearrangements, the three phages (*Staphylococcus* phage K, bacteriophage Felix 01, and bacteriophage P100) formed a clade that branched after the Mycoplasmatales. The difference in the three rearranged topologies is the branching order of the three phage sequences. Topological rearrangements were created in Treeview1.5 *(97)*. The likelihoods of each site and the global likelihoods for each tree were estimated using PUZZLE5.2 *(92)*, option-wsl with a WAG $+\Gamma$ (eight rate categories) $+$ I model of evolution. The likelihood values were used as input for CONSEL *(96)* to perform the AU test. All rearrangements have an associated P-value < 0.05 and therefore were rejected by the data, and are significantly worse alternatives to the original ML tree.

We have already seen how the LBA artifact can appear when multiple convergent changes along two fast evolving branches are interpreted as false synapomorphies by MP (**Fig. 9.8**). This is because too much homoplasy violates the assumption of minimum evolution inherent to maximum parsimony analysis. Simulation studies have shown MP is the method most sensitive to LBA artifact, whereas ML is the most robust, although not invulnerable *(64, 65, 67)*. As a result, most phylogeneticists consider ML as the most reliable method of phylogenetic inference *(68, 69)*.

In practice, LBA can be minimized by paying careful attention to taxon sampling. If long branches are clustered in a tree and an LBA artifact is suspected, there are several options that may alleviate the problem. First, taxa with long branches can be removed from the analysis, especially if they are unnecessary for addressing the specific evolutionary question. Where possible, taxon sampling can be modified so only slow evolving representatives of a taxonomic group are included *(70, 71)*. Another common option is to add species to the analysis that are closely related to the long branch in an effort to divide long branches as evenly as possible *(72, 73, 74)*. If LBA is suspected between fast evolving sequences of interest and the outgroup, then excluding the outgroup from the analyses facilitates testing this hypothesis. If the trees with and without the outgroup are significantly different, then the outgroup may be causing an LBA artifact.

3 Multigene Phylogenetic Analysis: An Example

A major topic in evolution is differentiating between groups of genes that share a common evolutionary history and genes that exhibit incongruent evolutionary histories. The amount of incongruence between markers is a crucial aspect to the debate surrounding the impact of LGT in prokaryotic evolution *(8, 9, 75, 76)*, but is also quite relevant to the study of phage evolution. The identification of groups of genes with congruent histories and the comparison between incongruent gene clusters may be particularly important in understanding the modularity of phage genomes *(15)*.

This subject is intimately linked to the amount of phylogenetic signal contained within markers. Genes with little phylogenetic signal will have low discriminative power between different trees. Comparison of their phylogenies will tell us little about their true evolutionary histories *(77)*. Phylogenetic trees inferred from genes with a robust phylogenetic signal, on the other hand, can be compared with the aim of uncovering the relative importance of vertical and lateral evolution in a group of biological entities.

A promising method called *Heat Map Analysis* (HMA) has been newly developed for assessing the strength of phylogenetic signal in markers (i.e., genes or proteins) and documenting incongruence between markers *(78)*. It involves the comparison of support/rejection patterns for multiple topologies and markers. For each marker, a list of *P*-values associated with a set of given topologies is generated, as in **Fig. 9.13**. The list of *P*-values is summarized into a matrix of genes and topologies and subsequently treated with a statistical clustering method *(79)*. HMA generates a graph through hierarchical clustering that allows the

simultaneous display of *P*-values for all combinations of genes and topologies together. As stated by Bapteste et al. (2005),

> genes that have the most similar responses to topologies, and topologies that are most similar in terms of the responses they evoke from genes, can be independently identified. More precisely, when applied to phylogenetics, "responses" are P-values for each set of genes, given those topologies. Clustering of genes allows identification of one or more sets of genes that might share a common evolutionary history. Clustering of topologies allows us to identify which trees are equally or nearly equally supported, and thus to assess how many distinct "best trees" there might be for a given dataset of genes.

HMA has been used to investigate the strength of phylogenetic signal and congruence between markers in several prokaryotic and a eukaryotic dataset *(78)*.

Recently, HMA was used to study the evolution of a collection of T4-type phages (Filee et al., personal communication). The purpose of this study was to identify genes that share a common evolutionary history and those with an incongruent history, which would shed light on the relative importance of inter-phage LGT in the evolution of the T4-type phages. The results of the study are presented here as an example of how the phylogenetic methods discussed previously can be applied to phage genomic data, and to illustrate the methodology behind HMA.

From a collection of 16 T4-type bacteriophage genomes, 24 core orthologs conserved enough for phylogenetic inference were identified (Filee et al., personal communication). These genes were largely located in two separate, syntenic regions of the phage genomes. One region contained a group of early-expressed genes involved in DNA replication and transcription, while the other encoded late-expressed genes, such as virion structural components. The majority of the 24 core genes appeared to be most closely related to other phage genes, however, the *nrdA* gene from *Aeromonas* phages was more similar to bacterial *nrdA* genes than those from other phage. Subsequent phylogenetic analysis demonstrated the *nrdA* gene in *Aeromonas* phage was acquired by LGT from a bacterium and therefore *nrdA* was removed from the set of core genes. The remaining 23 genes were concatenated (i.e., linked end-to-end) and a tree was inferred from the concatenated amino acid alignment by ML (**Fig. 9.15**).

Genes are often concatenated together into a large alignment in order to maximize the weak, but common, phylogenetic signal in the individual genes and to minimize random error. Gene concatenation has been used with varying levels of success to infer ancient relationship, where phylogenetic signal may

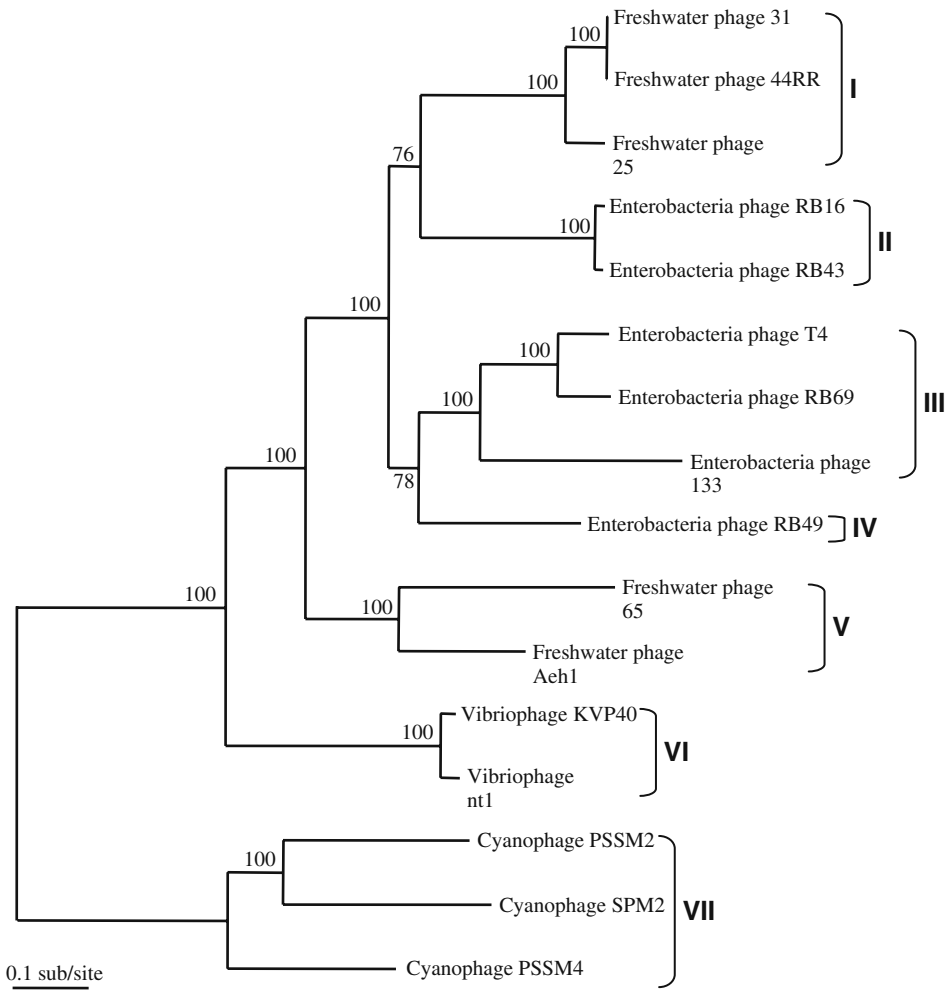


Fig. 9.15. Best ML tree inferred from a concatenation of 23 orthologous genes universally distributed in 16 T4-type phage genomes using PROML. The concatenation consisted of 4,653 amino acid sites. The evolutionary model was JTT $+\Gamma$ (four rate categories). The α value was estimated using PUZZLE5.2 *(92)* with global rearrangements and randomized input order (10 jumbles). PROML is part of the PHYLIP3.6 package *(89)*. Support for nodes corresponds to 100 pseudo-replicates estimated by the same method as the best ML tree. The seven phage groups that were rearranged to generate the 945 test trees employed in the HMA are labeled I to VII.

be particularly weak, such as between bacterial *(80)* and archaeal phyla *(81)*, or among eukaryotes *(82, 83)*. Although the concatenation method is being employed in this study of T4-like phage evolution, its general use in the study of phage evolution may not be appropriate. An assumption inherent to this methodology is that all genes in the concatenation share the same evolutionary history, which can be accurately reconstructed by overwhelming any phylogenetic noise, caused by homoplasy, with true phylogenetic signal (i.e., synapomorphy). In many cases, lateral gene transfer is viewed as phylogenetic noise, since it introduces homoplasy and disagreement into the dataset. However, for biological

forms such as phage, where LGT may be a frequent and important process, treating LGT as noise will not uncover the evolutionary processes shaping phage genomes. As alternatives, there are several single gene methods for uncovering genome mosaicism (i.e., LGT) that have been applied to the study of prokaryotic genomes that are applicable to the study of phage genome mosaicism. These include spectral *(30, 84)* synthesis *(85)* and splitstree analyses *(86)*.

With this caveat about gene concatenation in mind, the strength of the phylogenetic signal contained in the individual genes and the occurrence of LGT in the 23 core phage genes was investigated by HMA. Given each gene, a *P*-value was generated using the AU test *(56)* for the concatenation tree and an additional 945 rearrangements, resulting in 946 *P*-values for each of the 23 genes (21,758 *P*-values in total). The 945 additional trees correspond to the exhaustive set of topologies for seven groups (**Table 9.1**). The seven groups that were rearranged are labeled in **Fig. 9.15**. A heat map was generated by hierarchical clustering of *P*-values (**Fig. 9.16a**). In the heat map, the 946 trees are arrayed along the *y*-axis and the 23 genes are arrayed along the *x*-axis. Therefore, the heat map is a matrix and each entry in the matrix corresponds to the *P*-value for each gene/tree pair. Dark colors indicate a low *P*-value for a tree, given a gene (i.e., tree rejection). Light colors indicate a high *P*-value, given a gene (i.e., fail to reject tree). For the phage dataset, it is clear that most test trees are rejected by most genes, since the vast majority of trees receive a $P\text{-value} < 0.05$, given the vast majority of genes (**Fig. 9.16a**). Rejection of so many trees indicates most individual phage markers are highly discriminative between hypothesized trees and therefore contain a strong phylogenetic signal. For a visual comparison of a heat map where a set of genes fail to reject many trees see *(78)*.

In fact, there appears to be only a single test tree that is not rejected by the majority of the genes. Single rows are difficult to visualize in **Fig. 9.16a** because there are so many trees included in the heat map. However, the tree that is not rejected is displayed as the bottom row of **Fig. 9.16a**. That only one tree is not rejected by most markers indicates there is little incongruent signal between markers. The most likely explanation for a largely congruent signal for the phage core genes is that they share a common evolutionary history. If there was strong conflicting phylogenetic signal between markers, we would expect to see multiple trees, each supported by a subset of the markers. For example, it may have been the case that the genes within each distinct genomic region had a common evolutionary history, but, that the evolutionary history differed between the two genomic regions because these modules were shuffled between phage genomes. In this hypothetical case, the heat map would

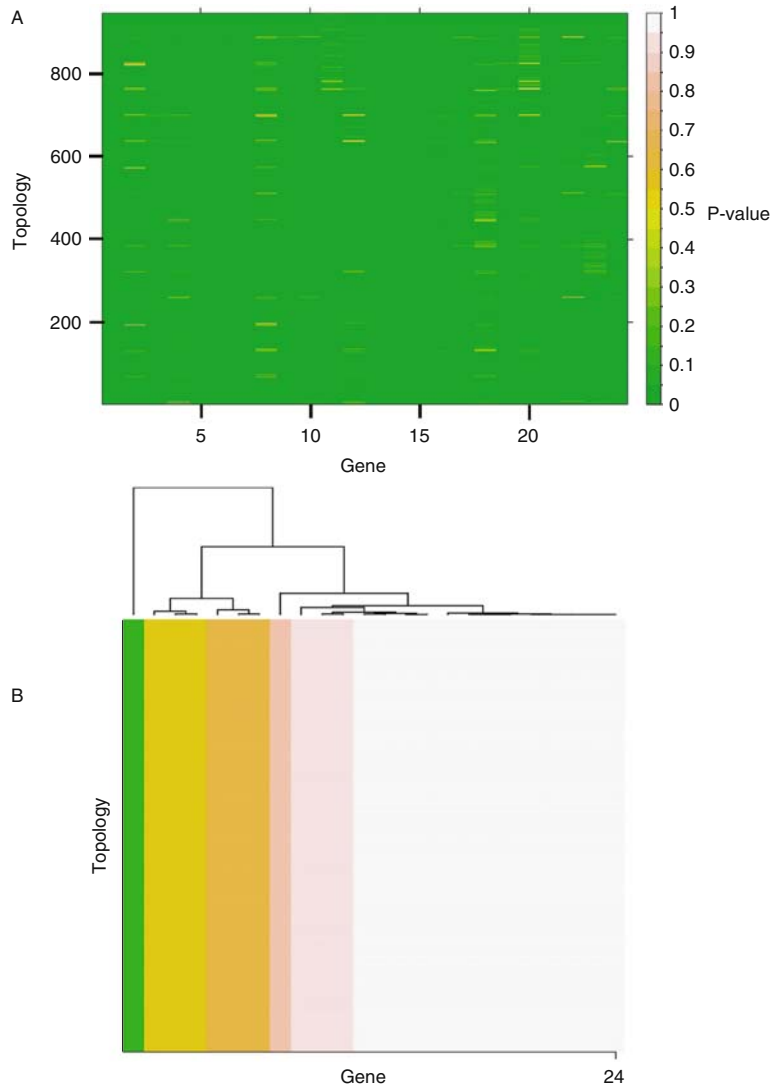


Fig. 9.16. (**A**) The heat map for the *P*-values generated for each of the 23 individual phage markers and the protein concatenation assessed against 946 test topologies. The 946 test topologies are arrayed along the *y*-axis and the 24 markers are arrayed along the *x*-axis. Each entry in the heat map matrix corresponds to a *P*-value for a single gene/tree pair. Dark colors indicate a low *P*-value (strong rejection of tree) and light colors indicate a high *P*-value (fail to reject tree). (**B**) The heat map for *P*-values of the 23 individual markers and the protein concatenation for the plausible topologies only (i.e., those trees for which the majority of the genes has a *P*-value >0.05), of which there was only one. The 24 markers were hierarchically clustered in relation to their support/rejection for this single tree. The dendogram of genes at the top of the heat map corresponds to the similarity of support/rejection values between genes. Only a single gene (*dark strip at the far left*) rejects this tree at $P=0.05$. For an in depth explanation of how a heat map is constructed, see *(78)*.

have revealed two non-rejected trees, where one would have been supported by the early-expressed phage genes and the other supported by late-expressed phage genes.

To further confirm the degree of congruence between phage markers, a heat map was constructed for the best trees (defined as those not rejected by the majority of the genes), which interestingly turned out to be the concatenation tree only. In this heat map, genes (columns) were clustered hierarchically with respect to their *P*-value for the concatenation tree (**Fig. 9.16b**). Therefore, genes with a similar pattern of support for the concatenation tree are clustered in the heat map. From this heat map, it becomes clear that only a single marker is incongruent with the rest of the dataset (dark column at the far left of **Fig. 9.16b**). This marker corresponds to gp13, which behaves atypically compared to most other genes in the dataset. Specifically, its individual phylogeny exhibits no basal support (Filee et al., personal communication) and there are probably few synapomorphies that support any single tree, including the concatenation tree, which results in a low likelihood.

Besides gp13, the heat map analysis suggests there is a core of genes within the T4-type bacteriophage that exhibit a similar evolutionary history, which most likely reflects vertical transmission. This is not to say LGT has not played a significant role in the evolution of T4-type bacteriophage. Outside of this core are many less ubiquitously distributed genes that appear to have been recruited from other sources by LGT (Filee et al., personal communication).

4 Concluding Remarks

The steady accumulation of phage genome sequences in public databases is providing a wealth of data amenable to study by molecular phylogenetics. Uncovering the degree of genetic mosaicism in different groups of phage will be of particular interest. In addition, documenting the patterns of LGT between phage and cellular life forms will provide insight into the evolution of the most diverse biological entities on Earth.

This chapter provides an introduction to the methods and models of evolution that should be considered when inferring a phylogeny from phage molecular sequence data. It remains to be determined how well these models of evolution based on cellular organisms accurately capture the true evolutionary process in phages. For example, empirical amino acid substitution matrices are all based on amino acid sequences from cellular organisms. Therefore, a little caution is warranted. From a selection of well-aligned molecular sequences, a tree should be inferred by several different methods in order to determine the effect of those

methods on hypothesized trees. Once a best tree is inferred, the robustness of the tree can be investigated by bootstrap analysis or comparing the best tree to other alternative trees. The effects of systematic error can be assessed by comparing trees constructed by different methods. For instance, if a relationship is well-supported by a distance method, but not by a ML method then it may not be a good estimate of the true evolutionary relationship. In the end, phylogenetics should be treated as a hypothesis-generating and hypothesis-testing process, since an inferred tree is only an estimation of the true evolutionary relationships modeled on current knowledge of molecular evolution.

Acknowledgements

We thank E. Bapteste and J. Leigh for their valuable input while preparing the manuscript and J. Filee for providing the HMA results prior to publication. We also acknowledge O. Zhaxybayeva, R.T. Papke, and J.E. Koenig for critical reading of the manuscript and W.F. Doolittle for providing a productive research environment. D.A.W. was supported by the Canadian Institute of Health Research and A.K.S. was supported by the Nova Scotia Health Research Foundation.

Appendix

Program and software packages available for performing phylogenetic analyses. Much of this information was taken from a website maintained by J. Felsenstein (http://evolution.genetics.washington.edu/phylip/software.html), which contains a wealth of information on phylogenetic programs. We recommend checking this website for program updates and new programs.

Program/Package (Author)	Applications	Availability
BioNJ (O. Gascuel)	An improved version of the NJ algorithm for inferring a tree from a distance matrix	Available at the website http://www.crt.umontreal.ca/~olivierg/bionj.html
CONSEL (H. Shimodaria)	Calculates the probability value to assess the confidence in the selection of phylogenetic trees. Includes AU, SH, and KH tests	Available at the website http://www.is.titech.ac.jp/~shimo/prog/consel/
ClustalX (D. Higgins, J. Thompson, T. Gibson, and F. Jeanmougin)	A widely used program for constructing multisequence DNA and protein alignments	Available at the website http://www.csc.fi/molbio/progs/clustalw/clustalw.html

(continued)

(continued)

Program/Package (Author)	Applications	Availability
fastDNAml (G. Olsen)	A fast method for inferring phylogeny from nucleotide sequences	Available at the website http://geta.life.uiuc.edu/~gary/programs/fastDNAml.html
FastME (R. Desper and O. Gascuel)	A fast algorithmic method for inferring a tree from a distance matrix	Available at the website http://atgc.lirmm.fr/fastme/
IQPNNI (L.S. Vinh and A. Haeseler)	Quartet puzzling and maximum likelihood program for inferring phylogenies from DNA and protein sequences	Available at the website http://www.bi.uni-duesseldorf.de/software/iqpnni/
MacClade (W.P. Maddison and D.R. Maddison)	A program with many capabilities for analyzing the evolution of discrete characters and molecular sequences. Widely used to visualize and manually edit multisequence alignments	Distributed by Sinauer Associates, Sunderland, Mass. 01375 USA Ordering information available at http://www.sinauer.com/
MEGA (S. Kumar, K. Tamaru, M. Nei)	A comprehensive programs that carries out maximum likelihood, parsimony, and distance methods for DNA and protein sequences	Available at the website http://www.megasoftware.net/
Modeltest (D. Posada and K. Crandell)	A program to test a hierarchy of models of DNA evolution by the Likelihood Ratio Test and the AIC	Available at the website http://darwin.uvigo.es/software/modeltest.html
MrBayes (J. Huelsenbeck and F. Ronquist)	A program for Bayesian inference of phylogenies from DNA and protein sequences	Available at the website http://mrbayes.csit.fsu.edu/
NJPlot (M. Gouy)	A program for plotting rooted trees	Available at the website http://pbil.univ-lyon1.fr/software/njplot.html
PAML (Z. Yang)	Maximum likelihood analysis of protein and DNA sequences, including analysis of codons. Model parameter estimation	Available at the website http://abacus.gene.ucl.ac.uk/software/paml.html

(continued)

(continued)

Program/Package (Author)	Applications	Availability
PAUP (D. Swofford)	A comprehensive package of programs that includes maximum likelihood, distance and parsimony analysis of DNA sequences. More information at http:// paup.csit.fsu.edu/	Distributed by Sinauer Associates, Sunderland, Mass. 01375 USA Ordering information available at http://www.sinauer.com/
PHYLIP (J. Felsenstein)	A comprehensive package of 35 programs, including distance, parsimony, maximum likelihood analysis of DNA and protein sequences	Available at the website http://evolution.gs. washington.edu/phylip.html
PHYML (S. Guindon, O. Gascuel)	A fast maximum likelihood program for nucleotide or protein sequences. Good for analyzing large datasets	Available at the website http://atgc.lirmm.fr/phyml/
PUZZLE (H.A. Schmidt, K. Strimmer, A. Haeseler)	Maximum likelihood analysis of DNA and protein sequences by quartet puzzling, model parameter estimation, calculating site likelihoods	Available at the website http://www.tree-puzzle.de/ A helpful script for running bootstrap analysis called PUZZLEBOOT is available from A. Roger at http://hades.biochem.dal.ca/ Rogerlab/Software/software. html#puzzleboot
SplitsTree (D. Hudson)	Carries out split decomposition and spectral decomposition to visualize conflicting phylogenetic signal in molecular data	Available at the website: http://www-ab.informatik. uni-tuebingen. de/software/ splits/welcome.html
TreeEdit (A. Rambaut and M. Charleston)	A program for organizing, manipulating and viewing sets of trees	Available at the website http://evolve.zoo.ox.ac.uk/ software.html?name=TreeEdit
Treeview (R. Page)	A program for displaying and manipulating trees	Available at the website http://taxonomy.zoology.gla. ac.uk/rod/treeview.html
Weighbor (W.J. Bruno, N.D. Socci, A.L Halpern)	A weighted version of the NJ algorithm that gives significantly less weight to the larger distances	Available at the website http://www.t10.lanl.gov/ billb/weighbor/index.html

References

1. Wilhelm, S.W. and C.A. Suttle. 1999. Viruses and nutrient sycles in the sea. Bioscience *49*:781–788.
2. Edwards, R.A. and F. Rohwer. 2005. Viral metagenomics. Nat Rev Microbiol *3*: 504–510.
3. Tetart, F., C. Desplats, M. Kutateladze, C. Monod, H.W. Ackermann, and H.M. Krisch. 2001. Phylogeny of the major head and tail genes of the wide-ranging T4-type bacteriophages. J Bacteriol *183*:358–366.
4. Filee, J., F. Tetart, C.A. Suttle, and H.M. Krisch. 2005. Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. Proc Natl Acad Sci U S A *102*:12471–12476.
5. Short, C.M. and C.A. Suttle. 2005. Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. Appl Environ Microbiol *71*:480–486.
6. Lawrence, J.G., G.F. Hatfull, and R.W. Hendrix. 2002. Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. J Bacteriol *184*:4891–4905.
7. Rohwer, F. and R. Edwards. 2002. The Phage Proteomic Tree: a genome-based taxonomy for phage. J Bacteriol *184*:4529–4535.
8. Gogarten, J.P. and J.P. Townsend. 2005. Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol *3*: 679–687.
9. Boucher, Y., C.J. Douady, R.T. Papke, D.A. Walsh, M.E. Boudreau, C.L. Nesbo, R.J. Case, and W.F. Doolittle. 2003. Lateral gene transfer and the origins of prokaryotic groups. Annu Rev Genet *37*:283–328.
10. Feil, E.J. and B.G. Spratt. 2001. Recombination and the population structures of bacterial pathogens. Annu Rev Microbiol *55*: 561–590.
11. Hanage, W.P., C. Fraser, and B.G. Spratt. 2005. The impact of homologous recombination on the generation of diversity in bacteria. J Theor Biol.
12. Papke, R.T., J.E. Koenig, F. Rodriguez-Valera, and W.F. Doolittle. 2004. Frequent recombination in a saltern population of Halorubrum. Science *306*:1928–1929.
13. Casjens, S.R. 2005. Comparative genomics and evolution of the tailed-bacteriophages. Curr Opin Microbiol *8*:451–458.
14. Pedulla, M.L., M.E. Ford, J.M. Houtz, T. Karthikeyan, C. Wadsworth, J.A. Lewis, D. Jacobs-Sera, J. Falbo, et al. 2003. Origins of highly mosaic mycobacteriophage genomes. Cell *113*:171–182.
15. Hendrix, R.W. 2002. Bacteriophages: evolution of the majority. Theor Popul Biol *61*: 471–480.
16. Mann, N.H., A. Cook, A. Millard, S. Bailey, and M. Clokie. 2003. Marine ecosystems: bacterial photosynthesis genes in a virus. Nature *424*:741.
17. Lindell, D., M.B. Sullivan, Z.I. Johnson, A.C. Tolonen, F. Rohwer, and S.W. Chisholm. 2004. Transfer of photosynthesis genes to and from Prochlorococcus viruses. Proc Natl Acad Sci U S A *101*:11013–11018.
18. Zeidner, G., J.P. Bielawski, M. Shmoish, D.J. Scanlan, G. Sabehi, and O. Beja. 2005. Potential photosynthesis gene recombination between Prochlorococcus and Synechococcus via viral intermediates. Environ Microbiol *7*:1505–1513.
19. Orengo, C.A., D.T. Jones, and J.M. Thorntom. 2003. Bioinformatics: Genes, Proteins & Computers. BIOS Scientific Publishers Ltd., Oxford.
20. Schuler, G.D. 1998. Sequence Alignment and Database Searching, p. 145–171. *In* A.D. Baxevanis, and B.F.F. Ouellette (Eds.), Bioinformatics: A practical Guide to the Analysis of Genes and Proteins. John Wiley & Sons, Inc., New York.
21. Bapteste, E., H. Brinkmann, J.A. Lee, D.V. Moore, C.W. Sensen, P. Gordon, L. Durufle, T. Gaasterland, et al. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. Proc Natl Acad Sci U S A *99*: 1414–1419.
22. Walsh, D.A., E. Bapteste, M. Kamekura, and W.F. Doolittle. 2004. Evolution of the RNA polymerase B′ subunit gene (rpoB′) in Halobacteriales: a complementary molecular marker to the SSU rRNA gene. Mol Biol Evol *21*:2340–2351.
23. Walsh, D.A., R.T. Papke, and W.F. Doolittle. 2005. Archaeal diversity along a soil salinity gradient prone to disturbance. Environ Microbiol *7*:1655–1666.
24. Huelsenbeck, J.P., B. Larget, R.E. Miller, and F. Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst Biol *51*:673–688.
25. Yang, Z. 1996. Phylogenetic analysis using parsimony and likelihood methods. J Mol Evol *42*:294–307.
26. Spencer, M., E. Susko, and A.J. Roger. 2005. Likelihood, parsimony, and heterogeneous evolution. Mol Biol Evol *22*: 1161–1164.

27. Swofford, D.L., G.J. Olsen, P.J. Waddell, and D.M. Hillis. 1996. Phylogenetic Inference, p. 407–514. *In* D.M. Hillis, C. Moritz, and M. B.K. (Eds.), Molecular Systematics. Sinauer Associates Inc., Sunderland.
28. Sokal, R.R. and C.D. Michener. 1958. A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin *38*:1409–1438.
29. Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol *4*:406–425.
30. Page, R.D.M. and E.C. Holmes. 1998. Molecular Evolution: A Phylogenetic Approach. Blackwell Science Inc., Malden.
31. Gascuel, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol *14*:685–695.
32. Bruno, W.J., N.D. Socci, and A.L. Halpern. 2000. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. Mol Biol Evol *17*:189–197.
33. Hollich, V., L. Milchert, L. Arvestad, and E.L. Sonnhammer. 2005. Assessment of Protein Distance Measures and Tree-Building Methods for Phylogenetic Tree Reconstruction. Mol Biol Evol *22*:2257–2264.
34. Rzhetsky, A. and M. Nei. 1992. Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. J Mol Evol *35*:367–375.
35. Nei, M. and S. Kumar. 2000. Molecular Evolution and Phylogenetics. Oxford University Press, Inc., Oxford.
36. Fitch, W.M. and E. Margoliash. 1967. Construction of phylogenetic trees. Science *155*:279–284.
37. Felsenstein, J. 2004. Inferring Phylogenies. Sinauer Associates, Inc, Sunderland.
38. Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Systematic Zoology *27*:401–410.
39. Kishino, H., T. Miyata, and M. Hasegawa. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J Mol Evol *31*:151–160.
40. Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol *17*:368–376.
41. Miller, E.S., J.F. Heidelberg, J.A. Eisen, W.C. Nelson, A.S. Durkin, A. Ciecko, T.V. Feldblyum, O. White, et al. 2003. Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage. J Bacteriol *185*:5220–5233.
42. Susko, E., Y. Inagaki, and A.J. Roger. 2004. On inconsistency of the neighbor-joining, least squares, and minimum evolution estimation when substitution processes are incorrectly modeled. Mol Biol Evol *21*:1629–1642.
43. Lio, P. and N. Goldman. 1998. Models of molecular evolution and phylogeny. Genome Res *8*:1233–1244.
44. Posada, D. and K.A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics *14*:817–818.
45. Dayhoff, M.O., R.M. Schwartz, and B.C. Orcutt. 1978. A model of evolutionary change in proteins., p. 345–358. *In* M.O. Dayhoff (Ed.), Atlas of Protein Sequence and Structure 5. National Biomedical Research Foundation, Washington.
46. Jones, D.T., W.R. Taylor, and Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comp Appl Biosci *8*:275–282.
47. Veerassamy, S., A. Smith, and E.R. Tillier. 2003. A transition probability model for amino acid substitutions from blocks. J Comput Biol *10*:997–1010.
48. Whelan, S. and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol *18*:691–699.
49. Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol *10*:1396–1401.
50. Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol *39*:306–314.
51. Gribaldo, S. and H. Philippe. 2002. Ancient phylogenetic relationships. Theor Popul Biol *61*:391–408.
52. Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution *39*:783–791.
53. Kishino, H. and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J Mol Evol *29*:170–179.
54. Shimodaira, H. and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol *16*:1114–1116.
55. Goldman, N., J.P. Anderson, and A.G. Rodrigo. 2000. Likelihood-based tests of topologies in phylogenetics. Syst Biol *49*: 652–670.

56. Strimmer, K. and A. Rambaut. 2002. Inferring confidence sets of possibly misspecified gene trees. Proc Biol Sci *269*:137–142.
57. Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst Biol *51*:492–508.
58. Vossbrinck, C.R., J.V. Maddox, S. Friedman, B.A. Debrunner-Vossbrinck, and C.R. Woese. 1987. Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes. Nature *326*:411–414.
59. Kamaishi, T., T. Hashimoto, Y. Nakamura, F. Nakamura, S. Murata, N. Okada, K. Okamoto, M. Shimizu, and M. Hasegawa. 1996. Protein phylogeny of translation elongation factor EF-1 alpha suggests microsporidians are extremely ancient eukaryotes. J Mol Evol *42*:257–263.
60. Keeling, P.J. and W.F. Doolittle. 1996. Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family. Mol Biol Evol *13*:1297–1305.
61. Keeling, P.J., M.A. Luker, and J.D. Palmer. 2000. Evidence from beta-tubulin phylogeny that microsporidia evolved from within the fungi. Mol Biol Evol *17*:23–31.
62. Inagaki, Y., E. Susko, N.M. Fast, and A.J. Roger. 2004. Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF-1alpha phylogenies. Mol Biol Evol *21*: 1340–1349.
63. Brinkmann, H., M. van der Giezen, Y. Zhou, G. Poncelin de Raucourt, and H. Philippe. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol *54*: 743–757.
64. Huelsenbeck, J.P. 1995. The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. Mol Biol Evol *12*:843–849.
65. Kuhner, M.K. and J. Felsenstein. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol *11*:459–468.
66. Gaut, B.S. and P.O. Lewis. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. Mol Biol Evol *12*: 152–162.
67. Swofford, D.L., P.J. Waddell, J.P. Huelsenbeck, P.G. Foster, P.O. Lewis, and J.S. Rogers. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst Biol *50*: 525–539.
68. Whelan, S., P. Lio, and N. Goldman. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet *17*:262–272.
69. Philippe, H., Y. Zhou, H. Brinkmann, N. Rodrigue, and F. Delsuc. 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol *5*:50.
70. Ruiz-Trillo, I., M. Riutort, D.T. Littlewood, E.A. Herniou, and J. Baguna. 1999. Acoel flatworms: earliest extant bilaterian Metazoans, not members of Platyhelminthes. Science *283*:1919–1923.
71. Ruiz-Trillo, I., J. Paps, M. Loukota, C. Ribera, U. Jondelius, J. Baguna, and M. Riutort. 2002. A phylogenetic analysis of myosin heavy chain type II sequences corroborates that Acoela and Nemertodermatida are basal bilaterians. Proc Natl Acad Sci U S A *99*:11246–11251.
72. Dacks, J.B., J.D. Silberman, A.G. Simpson, S. Moriya, T. Kudo, M. Ohkuma, and R.J. Redfield. 2001. Oxymonads are closely related to the excavate taxon Trimastix. Mol Biol Evol *18*:1034–1044.
73. Anderson, F.E. and D.L. Swofford. 2004. Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA. Mol Phylogenet Evol *33*:440–451.
74. Lin, Y.H., P.A. McLenachan, A.R. Gore, M.J. Phillips, R. Ota, M.D. Hendy, and D. Penny. 2002. Four new mitochondrial genomes and the increased stability of evolutionary trees of mammals from improved taxon sampling. Mol Biol Evol *19*:2060–2070.
75. Beiko, R.G., T.J. Harlow, and M.A. Ragan. 2005. Highways of gene sharing in prokaryotes. Proc Natl Acad Sci U S A *102*:14332–14337.
76. Kunin, V., L. Goldovsky, N. Darzentas, and C.A. Ouzounis. 2005. The net of life: reconstructing the microbial phylogenetic network. Genome Res *15*:954–959.
77. Bapteste, E., Y. Boucher, J. Leigh, and W.F. Doolittle. 2004. Phylogenetic reconstruction and lateral gene transfer. Trends Microbiol *12*:406–411.
78. Bapteste, E., E. Susko, J. Leigh, D. MacLeod, R.L. Charlebois, and W.F. Doolittle. 2005. Do orthologous gene phylogenies really support tree-thinking? BMC Evol Biol *5*:33.
79. Gordon, A. 1999. Classification.
80. Brochier, C., E. Bapteste, D. Moreira, and H. Philippe. 2002. Eubacterial phylogeny based on translational apparatus proteins. Trends Genet *18*:1–5.
81. Brochier, C., P. Forterre, and S. Gribaldo. 2005. An emerging phylogenetic core of Archaea: phylogenies of transcription and translation machineries converge following

addition of new genome sequences. BMC Evol Biol *5*:36.
82. Moreira, D., H. Le Guyader, and H. Philippe. 2000. The origin of red algae and the evolution of chloroplasts. Nature *405*:69–72.
83. Baldauf, S.L., A.J. Roger, I. Wenk-Siefert, and W.F. Doolittle. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science *290*:972–977.
84. Zhaxybayeva, O., P. Lapierre, and J.P. Gogarten. 2004. Genome mosaicism and organismal lineages. Trends Genet *20*:254–260.
85. MacLeod, D., R.L. Charlebois, F. Doolittle, and E. Bapteste. 2005. Deduction of probable events of lateral gene transfer through comparison of phylogenetic trees by recursive consolidation and rearrangement. BMC Evol Biol *5*:27.
86. Hudson, D.H. and D. Bryant. 2005. Applications of Phylogenetic Networks in Evolutionary Studies. Mol Biol Evol *On line early*.
87. Thompson, J.D., T.J. Gibson, F. Plewniak, F. Jeanmougin, and D.G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res *25*:4876–4882.
88. Maddison, D.R. and W.P. Maddison. 2003. MacClade 4. Sinauer Associates, Inc., Sunderland.
89. Felsenstein, J. 2004. PHYLIP: Phylogeny Inference Package Version 3.6.
90. Graur, D. and W. Li. 2000. Fundamentals of Molecular Evolution. Sinauer Associates Inc., Sunderland.
91. Guindon, S., F. Lethiec, P. Duroux, and O. Gascuel. 2005. PHYML Online – a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res *33*: W557–559.
92. Schmidt, H.A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics *18*:502–504.
93. Swofford, D.L. 1998. PAUP*: phylogenetic analysis using parsimony. Sinauer Associates, Inc., Sunderland.
94. Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci *13*:555–556.
95. Adachi, J. and M. Hasegawa. 1996. MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. Computer Science Monographs *28*.
96. Shimodaira, H. and M. Hasegawa. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics *17*:1246–1247.
97. Page, R.D.M. 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. Comp Appl Biosci *12*: 357–358.

Section II

Bacteriophage Transcriptomics and Proteomics

Chapter 10

Preparation of RNA from Bacteria Infected with Bacteriophages: A Case Study from the Marine Unicellular *Synechococcus* sp. WH7803 Infected by Phage S-PM2

Jinyu Shan and Martha Clokie

Abstract

Bacteriophages manipulate bacterial gene expression in order to express their own genes or influence bacterial metabolism. Gene expression can be studied using real-time PCR or microarrays. Either technique requires the prior isolation of high quality RNA uncontaminated by the presence of genomic DNA. We outline the considerations necessary when working with bacteriophage infected bacterial cells. We also give an example of a protocol for extraction and quantification of high quality RNA from infected bacterial cells, using the marine cyanobacterium WH7803 and the phage S-PM2 as a case study. This protocol can be modified to extract RNA from the host/bacteriophage of interest.

Key words: RNA, RNA extraction, RNA quantification.

1 Introduction

In bacteriophage biology, it is often important to quantify the expression of either phage or host genes in infected cells. The two main approaches to this are using real-time reverse transcriptase PCR (**Chapter 34**) or microarray analysis (**Chapter 35**). In approach, the quality and quantity of RNA template is critical. For real-time PCR, RNA extraction is followed by cDNA synthesis, and for microarray analysis amplification may be necessary.

Several methods can be used to extract RNA from bacterial cells *(1)*. RNA must be released from the cells which can be done by snap freezing and thawing, using enzymatic lysis, sonication or bead beating. RNA can then be extracted using organic solvents (in general, either hot phenol or guanidine

Martha R. J. Clokie, Andrew M. Kropinski (eds.), *Bacteriophages: Methods and Protocols, Volume 2: Molecular and Applied Aspects, vol. 502*, © 2009 Humana Press, a part of Springer Science+Business Media
DOI 10.1007/978-1-60327-565-1_10 Springerprotocols.com

thiocyanate–phenol/chloroform *(2)*, or using a membrane-based system). These methods are used to extract total RNA although there are also methods such as the Triton X-100 boiling method which can selectively remove ribosomal RNA and thus enhance the signal from the remaining mRNA *(1)*. It is beyond the scope of this chapter to exhaustively review all methods of RNA extraction.

In principal, RNA extraction from infected bacterial cells should not be particularly different than extraction from uninfected cells. However, there are a few important considerations to be aware of when extracting RNA from infected bacterial cells, and it is the purpose of this chapter to raise awareness of these issues. We also present a case study for the extraction of RNA from infected marine *Synechococcus* cells. The method presented uses a TRIzol extraction, two DNase I digestion steps and two on column purifications.

It is likely that well-studied bacteria will have well established and optimized methods of RNA extraction. However, when working with bacteriophages which infect less well-studied bacteria, it may be advantageous to compare several techniques to see which gives higher yields and quality of RNA. For example, in infected *Synechococcus*, we found the TRIzol reagent (a phenol and guanidine isothiocyanate solution made by Invitrogen) extraction method was more effective than that based on phenol/chloroform, both in terms of yield and quality of RNA.

Depending on the phage/bacteria of interest and the infection conditions, the infection experiment may run for a number of minutes to several days or weeks. Regardless of timescale, it is best to complete the experiment, and then perform the RNA extraction, once all the samples have been collected. In gene expression work, it can be difficult to ensure that one is comparing like with like, and to maximize consistency; all samples to be compared in a gene expression study should be extracted at the same time, and under the same conditions *(3)*. It is not possible to accurately estimate the concentration of cDNA, therefore RNA must be quantified as accurately as possible and the quality should be ascertained.

One of the most important hurdles in the quantification of gene expression is the complete removal of any contaminating DNA from the RNA samples. One particular problem which was very apparent in our system was how difficult it was to remove phage DNA, from the later time points of an infection cycle. There is more phage DNA present at later time points and furthermore, it is likely that phages modify their DNA using processes such as methylation *(4)*, so it is more difficult to remove than cellular DNA. Before the RNA can be used in further experiments, it is necessary to repeatedly test the extracted RNA to see if phage or host genes can be amplified from it using PCR

(**Chapter 26**). Only when no product is obtained can downstream application, such as cDNA synthesis be performed.

2 Materials

1. Exponentially growing *Synechococcus* cells.
2. Phage S-PM2 stock.
3. TRIzol Reagent (Invitrogen).
4. Chloroform (Sigma).
5. Bench-top microcentrifuge.
6. Isoprophyl alcohol (Sigma).
7. 75% ethanol (in nuclei acid-free water).
8. RNeasy Mini kit (Qiagen).
9. RNase-free DNase Set (Qiagen).
10. Nuclei acid-free water (Ambion).
11. β-Mercaptoethanol (Sigma).
12. DNase I and ×10 DNase I buffer (Ambion).
13. Water bath at 37 °C.
14. Nanodrop spectrophotometer (or other accurate spectrophotometer).
15. Agilent 2100 bioanalyzer (optional).

3 Methods

3.1 RNA Isolation

3.1.1 Sample Collection

1. Infect exponentially growing *Synechococcus* WH7803 ($OD_{750} = 0.35$) with phage S-PM2, under the required culture conditions and MOI (*see* Steve's chapters for more info here).
2. Collect samples (50 ml) at different time points during infection (*see* **Note 1**). Centrifuge samples at $4,000g$ for 10 min to pellet cells.
3. Snap freeze pellet in liquid nitrogen and store at −20 °C until RNA extraction.

3.1.2 RNA Extraction

1. Thoroughly re-suspend the cell pellet in TRIzol Reagent (1.5 ml).
2. Incubate at room temperature (20–25 °C) for 10 min.
3. Add $CHCl_3$ to a final concentration of 0.2 % and shake vigorously for 30 s.
4. Incubate at room temperature (20–25 °C) for a further 10 min.
5. Centrifuge at $15,000g$ in bench-top microcentrifuge for 10 min.
6. Transfer the upper (aqueous) phase to a new tube.

7. Add 0.5 volume isopropyl alcohol.
8. Leave on ice for 12–15 min.
9. Centrifuge at 15,000*g* in bench-top microcentrifuge for 10 min at 4 °C.
10. Wash the pellet with 75% ethanol followed by a further centrifugation at 15,000*g* in bench-top microcentrifuge for 10 min at 4 °C.
11. Remove the supernatant and dry the pellet in air for a maximum of 15 min (never let the sample dry out completely).
12. Dissolve the pellet in 90 μl nucleic acid–free water and 10 μl of ×10 DNase Buffer with the addition of 8 units of DNase I.
13. Incubate at 37 °C for 20 min.
14. Continue the purification using an RNeasy Mini Kit on-column DNase digestion with the RNase-free DNase according to the manufacturer's instructions.
15. Apply another around Ambion DNase I digestion to the resulting RNA suspension in nuclei acid–free water (30–50 μl) for the purpose of eliminating any trace residual phage and host genomic DNA (repeat steps 15 and 16, *see* **Note 2**).
16. Purify the resulting RNA again using an RNeasy Mini Kit (Qiagen).

3.2 Quantification and Quality Assessment of RNA Using Spectrophotometeric Analysis

For quantitative analysis, 1 μl of the RNA sample is assayed using an accurate spectrophotometer (preferably a Nanodrop spectrophotometer) to obtain the concentration of RNA. The absorbance ratios at 260 nm/280 nm (to check for protein contamination) and 260 nm/230 nm (to check for organic solvent contamination) are measured. For good quality RNA, the 260/280 ratio should be between 1.8 and 2.1 and the 260/230 ratio above 1.8.

3.3 Quantification and Quality Assessment of RNA Using Elelctrophoretic and Microfluidic Analysis

The traditional way to check the quality of RNA is to run the sample on a denaturing agarose gel stained with ethidium bromide. Intact RNA will give two distinct 23S and 16S bands, whereas degraded RNA sample will have a smeared appearance. A ratio of roughly 2:1 (23S:16S) will confirm the integrity of RNA (although total RNA with lower rRNA ratios can also be of good quality (http://www.ambion.com/techlib/tn/111/8.html)). This is the case for the total RNA from *Synechococcus* sp. WH7803 where the 23S rRNA is unstable and thus displays as unfragmented 23S rRNA peak and two fragmented 23S-derived peaks (**Fig. 10.1**). This instability may be due to its size as well as its high degree of secondary and tertiary structure. An AU-rich sequence called a "hidden break" in

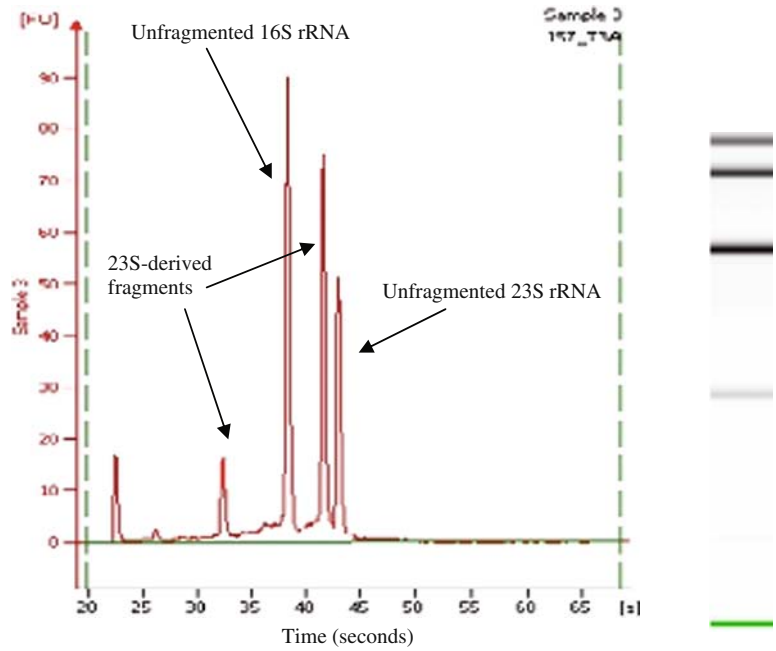


Fig. 10.1. Typical electropherogram and gel-like image of a high quality, *Synechococcus* sp. WH7803 total RNA sample. 23S rRNA is fragmented into two 23S-derived smaller rRNAs.

some 23S has been suggested that can result in processing of these rRNA species into two smaller RNAs.

If you are fortunate to have access to one, an Agilent 2100 bioanalyzer is a useful tool with which to check the quality of RNA *(5)*. This instrument uses microfluidics to provide detailed information about the condition of RNA samples. The output from the analysis shows an electropherogram image that provides a detailed visual inspection of RNA quality, it also converts this data to gel-like image of the sample and it gives the 23S/16S peak ratio, and approximate RNA concentration (*see* **Note 3**).

3.4 Ensuring Contaminating Phage DNA is Removed from Samples

It is essential to remove all traces of DNA from samples before proceeding with a cDNA synthesis or hybridization studies. It is extremely easy to not remove all DNA from RNA samples, however, this renders subsequent real-time PCR data sets meaningless. As described in the protocol, we removed DNA using both on column treatment and using DNase I. The best way to verify that an RNA sample is free from DNA is to run PCR experiments against RNA samples that have not reverse transcribed. If an RNA sample is contaminated with genomic DNA, a PCR product will be generated and further purification must be carried out.

3.5 Downstream Applications

This section is really to point out that there are no "bacteriophage-specific" considerations to be taken into account

when either making cDNA synthesis (for real-time PCR) or amplifying RNA for arrays. Standard protocols can be followed for these procedures.

4 Notes

(1) A 50 ml of cells yielded around 200–300 ng of RNA per microliters—using phenol/chloroform-based approach yielded less than 100 ng of RNA per microliters.

(2) These steps were necessary in our system to remove all contaminating phage genomic DNA. The earlier time points in the sample were easier to remove phage DNA from and so all contamination was removed without the final enzymatic treatment, but for consistency the same procedures must be performed on all samples. It will be necessary to prove that you have removed all DNA from the RNA sample before you proceed with the cDNA synthesis step. This can be done by performing PCR (**Chapter 26**) on the sample. If no DNA is present, there should be no amplification of genes from the sample. Both host and phage genes should be tested.

(3) Although the concentration is estimated using the bioanalyzer, we consistently found that the Nanodrop spectrophotometer gave a more reliable result.

References

1. Sung, K., et al., A simple and efficient Triton X-100 boiling and chloroform extraction method of RNA isolation from Gram-positive and Gram-negative bacteria. *FEMS Microbiology Letterss*, 2003. **229**(1): 97–101.
2. Chomczynski, P. and N. Sacchi, Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Analytical Biochemistry*, 1987. **162**(1): 156–159.
3. Bustin, S.A., Real-time, fluorescence-based quantitative PCR: a snapshot of current procedures and preferences. *Expert review of molecular diagnostics*, 2005. **5**(4): 493–498.
4. Casadesus, J., D. Low, and M.A.M.B.R.-S. 2006, Epigenetic gene regulation in the bacterial world. *Microbiology and molecular biology reviews*, 2006. **70**(3): 830–856.
5. Imbeaud, S., et al., Towards standardization of RNA quality assessment using user-independent classifiers of microcapillary electrophoresis traces. *Nucleic Acids Research*, 2005. **33**: e56 (Online publication).

Chapter 11

Quantification of Host and Phage mRNA Expression During Infection Using Real-Time PCR

Dr Martha R. J. Clokie

Abstract

Real-time, or quantitative PCR, is a valuable technique useful in bacteriophage research to quantify the abundance of phage or host gene transcripts. It can be used during the infection cycle both to monitor the expression of individual viral transcripts and to compare relative gene expression levels throughout the infection cycle. It is fairly economical to conduct and is useful in bacteria–phage systems where obtaining high yields of RNA is problematic. To perform real-time PCR, it is simply necessary to know the DNA sequence of the genes to be monitored, to have accurately quantified mRNA good quality cDNA, and access to a light-cycler. Although this chapter briefly reviews the basic principles of real-time PCR, the emphasis is on aspects of technique that are specific to the study of bacteriophage transcriptomics. These include (1) the selection of the target gene, (2) the choice of calibrator and reference genes, (3) RNA isolation for cDNA synthesis and (4) subsequent analysis of samples. This chapter should also be useful to those wishing to amplify genes from other types of templates such as metagenomic DNA or RNA extracted either from filtered samples or from agarose gels.

Key words: Real-time PCR, SYBR green, probes and primers, endogenous control, calibrator.

1 Introduction

Real-time or quantitative PCR combines PCR (**Volume 2 Chapter 2**) with spectrophotometry such that an increase in DNA concentration is measured in 'real-time' as opposed to standard PCR, where, the product size and approximate quantity is measured on an agarose gel at the end of the reaction. Unlike conventional PCR, the conditions of the reactions are standardized for all transcripts to allow multiple transcripts to be examined during each analysis. Real-time PCR works best when the region amplified (the amplicon) is small. In addition to the usual components for

Martha R. J. Clokie, Andrew M. Kropinski (eds.), *Bacteriophages: Methods and Protocols, Volume 2: Molecular and Applied Aspects, vol. 502*, © 2009 Humana Press, a part of Springer Science+Business Media
DOI 10.1007/978-1-60327-565-1_11 Springerprotocols.com

PCR, a fluorophore (either SYBR green or a fluorescent probe) is included in the reaction mix and the reactions are run in clear plastic so the amount of fluorescence can be measured during the relevant stage in the amplification procedure (after polymerization has been competed). The fluorescence level increases exponentially with the concentration of PCR product and is measured and plotted logarithmically against the cycle number. A threshold value (C_T) is then set so that the samples can be compared. This threshold value can be at any point in the exponential phase of the amplification (**Fig. 11.1**). To give an example of how C_T values allow a comparison of transcript abundance, at any point in the exponential phase, if a particular transcript has a C_T value of 15 at a particular threshold, then half as much template at that threshold would give a C_T value of 16, and half as much again would give a C_T value of 17 and so on.

Depending on the experimental design and plate layout, analysis of these data allows either the absolute or the relative amount of a gene transcript in a sample to be estimated. In absolute quantification, the absolute quantity of a target sequence from a single nucleic acid from a particular sample is determined, whereas in relative quantification the amount of target sequence is determined relative to the amount of the same (or different) target sequence in a calibrator sample. In bacteriophage transcriptomic

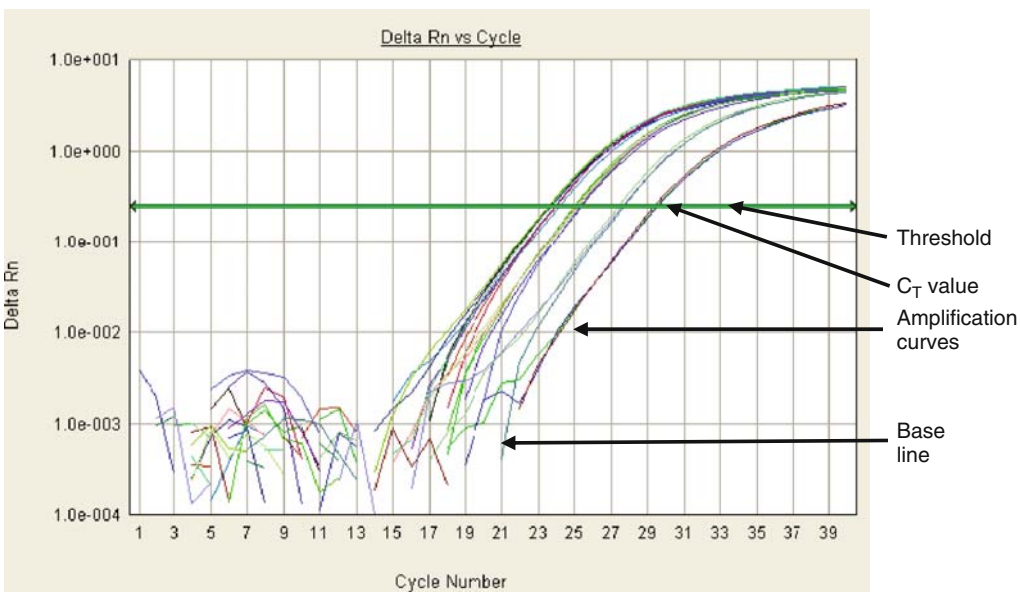


Fig. 11.1. Typical amplification curves from four samples of cDNA. Note amplification begins shortly above the baseline, and the C_T or threshold value is set during the exponential phase of the reaction. The point at which the C_T value is calculated is shown for the amplification curve at the far right of the curves. The *x*-axis shows the cycle number and the *y*-axis relative fluorescence and Delta Rn vs Cycle = Delta.

research, this calibrator is likely to be a particular point in a time course. Thus, for example, we may wish to determine how much transcript encoding the major capsid protein is present 10 minutes after infection compared to 1 minute after infection. While absolute quantification is useful in some scenarios, relative quantification is generally more useful in bacteriophage transcriptomics as it is often necessary to know how much of a transcript is present relative to a particular time point. Relative quantification has the advantage that accurate comparisons between samples can be determined over several plates without using standard curves on each plate.

As with real-time PCR in any system, the major decision to be made is between the technology based on SYBR green or which uses specific probes and primers *(1)*. SYBR green binds to double-stranded DNA and fluoresces when it is bound. There are many different designs of specific primers and probes, made by an equally large number of biotechnology companies (e.g. Applied Biosystems and Sigma). In general, the makers of specific real-time PCR machines recommend their own particular chemistries (though they are often not the cheapest), and most chemistries should be compliant with most machines. The probes used in real-time PCR anneal to the DNA in the region between the two primers, and in general they have two dyes, one which fluoresces and the other which is a quencher molecule. Fluorescence only occurs, and thus is detected, when the polymerase cleaves the fluorescent dye from the probe *(1)*. There are many variations on this theme.

The advantage of SYBR green is that it is much cheaper than using probes and primers and therefore numerous genes can be examined simply for the extra cost of additional primer pairs. It is also useful if the gene is variable in the region between the primers, as only the DNA which encodes the primer sequences that needs to be conserved. This is in comparison to using probes as well as primers, as the probe hybridizes to the region between the primers so also has to be conserved. Thus SYBR green could be used to quantify the amount of template present for a mixed population a gene where some versions have a mutation between the primers. The major disadvantage to SYBR green, however, is the lack of specificity, the greater risk of contamination and the longer time-period necessary to optimize the experiments. However, the specificity can be analysed using a melting curve (T_M) analysis following amplification. Here the finished plates containing the PCR products and SYBR green are heated to 95 °C. During this time, the instrument records the drop in SYBR green fluorescence that results from the dissociation of double-stranded DNA, primer dimers will have a lower melting temperature than that of the product, and more than one curve would represent multiple products.

The main advantage to a probe-based approach is its specificity, as the fluorescent signal is generated from the binding of the probe and target rather than simply to all double-stranded DNA. Less optimization is therefore required for the reaction and generally contamination is much less of a problem. A further advantage to probes and primers is that duplex experiments can be performed, where two products are amplified using different fluorescent probes (*see* **Note 1**). A disadvantage is the cost, however, when the cost of a researchers time is taken into account this may be negated.

The widespread infection of all bacteria with bacteriophages is increasingly apparent. It is often desirable to establish at what stage bacteriophages are in during their infection cycles. This important information can be obtained from an examination of the infected cell transcriptome. Real-time PCR is a highly sensitive tool with which to probe this in order to monitor the presence of transcripts either throughout a lytic infection cycle or during infection by a lysogen. It can be used independently to monitor transcript expression, or as a tool to corroborate and calibrate a micro-array data set *(2, 3)*.

Although bacteriophages must inject their DNA, replicate their genomes and construct protein coats, the order and timing of these events has only been established for a few phages [e.g. T4 *(2, 3)*]. Furthermore, expression profiles of the plethora of host and genes of unknown function which are found inside newly sequenced genomes are generally uncharted.

Real-time PCR is a standard technique which is increasingly used for quantification purposes. There is a large body of literature detailing developments on chemistries, experimental design and analysis. This was nicely summarized in a recent review by Bustin *(4)*. Useful introductions to the principles and applications of real-time PCR can be found in the literature provided by manufacturers' of real-time PCR machines (e.g. from Applied Biosystems, Corbett, and Stratagene). There is also a recent excellent and comprehensive book on the subject *(5)*. However, relatively little has been written specifically on the use of real-time PCR in bacteriophage–host relationships and this chapter represents the first methods overview on the subject. Real-time PCR can also be applied to examine the quantity of phage DNA from a range of templates. In which case, the protocol is even easier; simply follow the guidelines in **Chapters 23** and **Volume 2 Chapter 1** to prepare a high-quality DNA template, then all else applies. Other templates that are suitable for real-time PCR analysis include metagenomic samples (where, for example, the amount of a particular phage, or phage gene in a particular viral or bacterial DNA fraction could be determined *(6, 7)*).

2 Materials

2.1 Good Quality Template

The extraction of mRNA is covered in **Chapter 32**. Although real-time PCR can detect DNA concentrations as low as 1 fg, for routine real-time PCR, best results may be obtained when using between 1 and 100 ng of mRNA to be transcribed into cDNA. For each reaction well on the RT-PCR plate between 1 and 10 ng of cDNA will be used.

2.2 Primers and Probes

Primers can be ordered from your regular supplier (e.g. Invitrogen), and probe and primer combinations can be obtained from several biotech companies (e.g. Applied Biosystems). A variety of online resources can help you to design appropriate primers (*see* http://molbiol-tools.ca/PCR.htm). Most companies offer two routes to obtain probe and primer assays with either the option of designing primers oneself, or, they will design the assays for you. Often only the latter is guaranteed to work, but you have the option of selecting the approximate gene region in which you would like the probe to be placed, as it is possible to mask all areas of the gene that you do not wish primers/probes to hybridize. The products must be between 60 and 200 bp and the primers must anneal at 60 °C *(8)*. There are various software programmes which specifically design primers for real-time PCR (e.g. Primer Express from Applied Biosystems).

2.3 Master Mix

This contains the reagents that are present in standard PCR, including polymerase, buffer, $MgCl_2$, dNTPs and water and SYBR green (when used). This mix generally is at 2X strength and can be ordered from several biotech companies (e.g. Applied Biosystems, Qiagen, Sigma) (*see* **Note 2**).

2.4 Nuclease-Free Water (e.g. Sigma)

Can either be made in the laboratory using autoclaved filtered water or bought from companies such as sigma.

2.5 Plates and Adhesive Covers or Tubes

Either 96- or 386-well plates, if access to a robot is available. Adhesive covers must be heatproof and of optical quality. Some machines (e.g. Corbett) use individual tubes, and may take up to 72 individual tubes per reaction.

2.6 Light Cycling PCR Machine

Several companies now make light-cyclers that are based on a Peltier block in which a 96-well plate sits. These include Applied Biosystems, Qiagen and Eppendorf. The alternative is to have a rotary chamber in which the tubes sit and air is pumped in and out at the correct temperature (e.g. Corbett). The data acquisition is particularly fast in the latter systems and the uniformity of heat distribution is greater.

2.7 Bench Top Vortexer

2.8 Laminar Flow Hood and Related Precautions

It is important to minimize exposure of samples to nucleases and any form of contaminants. To avoid any problems from the outset, it is highly recommended that all real-time PCRs are setup in a laminar flow hood which has been sprayed with anti-nucleases. Gloves must be worn at all times, all plastic ware must be nuclease-free and pipette tips should have filter plugs.

3 Methods

3.1 Time Point Selection and Phage Infection

It is important to make sure that the time points chosen for a real-time PCR are representative of the different growth phases of the phage in question. To probe the transcriptome, one should have prior knowledge of the absorption rate of the phage. This will allow the infections to be synchronized and prevent the signal being muddled by the noise from second rounds of infection (**Chapter 16**). Other physical parameters should be measured such as the eclipse and latent periods (**Chapter 16**). Time points can therefore be taken at different phases of the cycle which will range from minutes (i.e. for T4 infecting *E. coli* to hours or even days, for example, in S-PM2 infecting *Synechococcus*). Procedures for phage infection and RNA isolation over a time course generally should follow the considerations described in **Chapter 10**. In particular, the RNA must be preserved rapidly and stabilized at each time point where reagents such as RNAlater® (Qiagen) can be extremely useful.

3.2 Target Gene Selection

To undertake real-time PCR, the sequence of the transcript of interest must be known. The transcript may either be part of a gene with known function, or possess homology to a gene with a known function. Alternatively its function may be unknown, and therefore its expression profile may at least show at what point during the infection cycle it is transcribed.

If the researcher is studying a new phage–bacteria system, then it may be useful to obtain information relating to genes where the expression profile is known in related phages. This may include structural genes or known 'early' or late genes. These data can give a framework of expression in which to compare the expression of novel genes. It may be that the researcher has a specific gene in whose function and expression profile they have an interest. However, it may be that the question of interest is more general. For example, the researcher may simply be interested in when phage capsid proteins are synthesized, in which case

it would be sensible to monitor the expression of the major capsid protein. For T4, the expression profiles of most genes has been determined *(2,9)* and proves a useful starting point when evaluating gene expression of under-studied T4-type phages which infect other bacteria.

3.3 Control Calibrator Gene Selection

One of the main difficulties with bacteriophage transcriptomics is that there is not necessarily one gene that is constitutively expressed throughout the infection cycle. For example, T4 switches off the replicative machinery of its host shortly after infection *(9)*. Being reliant on its host as an energy source, the cyanobacterial phage S-PM2 appears not to have this method of total host switch off *(10,11)*. As is the situation in other transcriptional studies, the ribosomal machinery appears to represent the best of an imperfect solution. Certainly, it seems that in the case of S-PM2, the amount of ribosomal RNA as represented by the 16S transcript remained constant at all time points. One problem with using the ribosomal mRNA as an endogenous control (calibrator) is that because it is present at a much greater abundance than any of the bacteriophage or host genes of interest, it needs to be analyzed using a different threshold sensitivity compared to all of the other genes. Other control genes commonly used include *rpoD, gyrA* and so on. None is perfect and each may present specific concerns depending on the phage–host system under study. Having available the sequence for a candidate calibrator gene from the bacterial host will often determine what is selected. The highly conserved regions of 16S rRNA typically circumvent this problem. When designing primers for the calibrator, one should follow the same considerations used for the target gene primers.

One way to circumvent the problem of identifying a gene with which to normalize data is to normalize to the total amount of RNA extracted from the cells. Unfortunately, this approach also has its pitfalls as it is difficult to accurately quantify the cDNA, and discrepancies may arise between samples collected at different time points and unequal conversion of mRNA to cDNA. Although these are valid concerns, I recently analyzed a data set consisting of the expression of 12 bacteriophage genes and 5 host genes during an infection and obtained very similar results regardless of whether I normalized to 16S or to total mRNA (unpublished). Thus although normalizing to either 16S mRNA or total mRNA is not perfect, it is reassuring that they essentially report the same information.

The choice of calibrator is always difficult, and not a problem specific to phage–bacterial work, the researcher should probably start by using 16S as a calibrator, and if possible test the levels of expression of host genes that they believe may not be influenced by phage infection.

3.4 Primer or Primer and Probe Optimization

It is important to test all primers or probes and primer combinations before analyzing the valuable cDNA samples obtained from the experiments. One must ensure that amplification is equally efficient and reports correctly across a range of template concentrations. If the researcher is using relative quantification, then after the standard curves have been carried out once, it is not necessary to repeat them on each plate.

For each gene to be assayed, make a serial dilution of either cDNA or phage genomic DNA, so samples contain 10, 5, 2.5, 1.25, 0.625 and 0.3125 ng of template. Add the appropriate amount of probe/primer and master mix. Run the reaction and check that the slopes of the amplification curves are parallel in log view and that for each sample that reduced by half in the quantity of cDNA, there is an increase of 1 in the C_T value.

If the complete genome sequence of the bacteriophage is known, and if it has one copy of all the genes for which you wish to monitor the expression, then it serves as a natural template with which to compare the efficiency of amplification of gene transcripts, and to establish a standard curve. If the genomic template is amplified equally efficiently for each gene primer set, then the threshold value should be identical for each gene. One then assumes that equal efficiency will occur with the cDNA as template, although due to differences in transcript abundance, this cannot always be confirmed.

3.5 Plate Design

Having established that the primers and probes work and report evenly for all target and calibrator genes, one can then proceed to the experiment. **Table 11.1** shows an example of a typical setup for the relative quantification of a number of bacteriophage genes using real-time PCR. All assays are performed in triplicate and should contain a negative control for each transcript. The negative controls should contain everything but template. If using relative quantification then fluorescent controls should also be present to allow cross-plate comparisons. Thus the setup on this plate would allow a regular relative quantification, cross-plate comparisons and an estimation of the actual number of bacteriophage transcripts (see below).

The experiment shown in **Table 11.1** had five time points; 0, 1, 3, 6 and 9 hours after infection. This covered the latent period for the phage S-PM2 infecting *Synechococcus* WH7803 at an MOI of 10. The expression profiles were monitored for one bacteriophage gene of interest (D1p) and three host genes (D1h, D1I1 and D1I2) and the ribosomal RNA (16S) was used as an endogenous control.

The plate also dedicates 12 wells to fluorescent controls. These controls consist of four different bacteriophage genes which previously showed equal amplification from purified

Table 11.1
A typical RT-PCR plate setup. The numbers 0, 1, 3, 6 and 9 represent time after infection. D1h is an assay that detects both alleles of host *psbA*, D1I1 and D1I2 only detect one allele, D1p is an assay that detects bacteriophage-encoded DNA, 16S is an endogenous control, RegA, gene 44, 47 and 18 are both fluorescent controls and calibrators. An endogenous control is present for each sample point and a negative for each assay

	1	2	3	4	5	6	7	8	9	10	11	12
A	0 A	0 A	0 A	1 A	1 A	1 A	3 A	3 A	3 A	6 A	6 A	6 A
	D1h	D1h	D1h	D1h	D1h	D1h	D1h	D1h	D1h	D1h	D1h	D1h
B	9 A	9 A	9 A	0 A	0 A	0 A	1 A	1 A	1 A	3 A	3 A	3 A
	D1h	D1h	D1h	D1p	D1p	D1p	D1p	D1p	D1p	D1p	D1p	D1p
C	6 A	6 A	6 A	9 A	9 A	9 A	0 A	0 A	0 A	1 A	1 A	1 A
	D1p	D1p	D1p	D1p	D1p	D1p	D1I1	D1I1	D1I1	D1I1	D1I1	D1I1
D	3 A	3 A	3 A	6 A	6 A	6 A	9 A	9 A	9 A	0 A	0 A	0 A
	D1h	D1h	D1h	D1h	D1h	D1h	D1h	D1h	D1h	D1I2	D1I2	D1I2
E	1 A	1 A	1 A	3 A	3 A	3 A	6 A	6 A	6 A	9 A	9 A	9 A
	D1I2	D1I2	D1I2	D1I2	D1I2	D1I2	D1I2	D1I2	D1I2	D1I2	D1I2	D1I2
F	0 A	0 A	0 A	1 A	1 A	1 A	3 A	3 A	3 A	6 A	6 A	6 A
	16S	16S	16S	16S	16S	16S	16S	16S	16S	16S	16S	16S
G	RegA	RegA	RegA	44	44	44	47	47	47	18	18	18
	10 ng	10 ng	10 ng	10 ng	10 ng	10 ng	10 ng	10 ng	10 ng	10 ng	10 ng	10 ng
	DNA	DNA	DNA	DNA	DNA	DNA	DNA	DNA	DNA	DNA	DNA	DNA
H	9 A	9 A	9 A	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg
	16S	16S	16S	RegA	44	47	18	D1p	D1h	D1I1	D1I2	16S

S-PM2 DNA. Their presence here is to allow easy cross-plate comparisons as 10 ng of purified bacteriophage DNA from the same stock was used for each fluorescent control. As these measurements do not vary, the threshold at which the C_T value is chosen can be maintained throughout all plate comparisons. When further host and phage genes are examined from these samples, the C_T values will be comparative to each other, by checking the C_T values of these fluorescent genomic DNA controls.

These fluorescent controls have a second bacteriophage-specific purpose as calibrators. Because it is possible to calculate how many phage genomes must be present in a known

concentration of DNA (*see* **Note 3**), the number of transcripts present in the expression work can be calculated by comparison to the genomic viral DNA samples.

One final point regarding fluorescent controls is that the use of an assay (e.g. primers and probes for gene 47) as a fluorescent control on gDNA template, does not preclude its use on cDNA from infection experiments.

3.6 Plate Setup

A standard total volume for each reaction is 20 μl. To minimize costs, and when a good procedure has been established, this may be reduced further to half volumes or less. Each reaction contains:

1) 2X reaction buffer (10 μl).
2) 20X Gene expression mix (1 μl expression mix or 1 μl each of different primers and 1 μl of SYBR green).
3) Template (variable)
   i) cDNA (10 ng in 7–9 μl water)
   ii) genomic DNA (in 7–9 μl water)
   iii) negative control (water only)
4) Nuclease-free water to total volume 20 μl.

One way to setup a large plate with several assays is to make two sets of master mixes. First create a probe/primer or primer/SYBR green master mix (total volume 11 or 13 μl, respectively, depending on whether SYBR green or primer/probes are being used). Add this to all of the wells/tubes and then add the template, water and mix, where template is at correct concentration (e.g. 10 ng in 9 or 7 μl of water). In other commercial mixes, the Master Mix solution is at 2X with all the reagents, including SYBR green, in which case 10 μl of mix is added to the template which contains 1–10 ng of cDNA, the primers and water to a total volume of 20 μl.

3.7 Running the Reaction

Analysis will be easier if the plate is setup correctly. Although the details vary from machine to machine, the following commonalities apply. Each transcript assayed should be assigned an individual detector. Essentially, even if FAM or SYBR green is being used for each transcript, you designate them separate 'detectors' as this allows separate analysis of each transcript. This is particularly useful, for example, if the 16S needs to be analyzed using different settings. If you are performing relative quantification, the endogenous controls should be specified and if the experiment has multiple samples then there should be an endogenous control for each sample.

Cycle conditions are typically standardized, with fluorescent measurements taken at the end of each cycle. If using SYBR green perform a melting curve analysis at the end of the cycling to show if more than one product has been amplified.

3.8 Analysing the Data

The first step of any real-time PCR analysis is to adjust the baseline and threshold value (**Fig. 11.1**). Generally, the automatically generated baseline will be optimal but if not it should be adjusted as necessary (e.g. if plates are to be compared and the standards vary between plates that contain identical samples). The baseline should be set such that amplification occurs just after the maximum baseline value for all of the samples (*see* **Fig. 11.1**). The threshold or C_T values should be adjusted such that it occurs in the early or middle exponential phase of the amplification curve. When these have been decided, it is necessary to reanalyze the data. Any C_T values which are obviously in error can be removed from the analysis (for example, if two C_T values are 15 and 1 is 45—one could trust the two readings which are in agreement). After an appropriate baseline and C_T value have been chosen, the results can be analysed and exported either to a Microsoft Excel spread sheet or to a custom real-time analysis software (often this is software unique to the light-cycler).

There are many ways in which to analyze both absolute and relative real-time PCR data and excellent help with dozens of useful references can be found on a website written and maintained by Michael Pfaffl (http://www.gene-quantification.info/). It is outside the scope of this chapter to exhaustively review analysis approaches, but for reviews on approaches to data analysis, *see (4, 12, 13)*.

Although the method of analysis should be decided upon before designing the experiment, if uncertain it is often best to add appropriate controls so as not to exclude future analysis possibilities. The setup shown in **Table 11.1** contains enough references within the 96-well plate to be analyzed in several ways. Thus it could be normalized to the 16S using the $2^{\Delta\Delta CT}$ method (see below) or analysed using a modification of the method detailed below. Rather than exhaustively review analysis approaches, a specific example of analysis is given which I have found to be particularly useful. It is based on a modification of the $2^{\Delta\Delta CT}$ method and allows data to be presented in terms of number of phage transcripts.

In the standard $2^{\Delta\Delta CT}$ method, the amount of transcript in a sample is compared to an endogenous control and then to a calibrator *(14)*. For this calculation to be valid, the amplification efficiency of the target (gene of interest) and the calibrator must be approximately the same. This can be determined by performing serial dilutions of template for both genes as described in **Section 3.3**. The C_T values at each dilution can then be subtracted from each other as a ΔC_T for each concentration and plotted on the *y*-axis against the log of the input amount of total genomic DNA. The slope of the graph should be < 0.1 *(14)*. Once even amplification efficiency has been established, at

all time points, the C_T value of the endogenous control (in this case 16S at each time point) is subtracted from the time point to give ΔC_T and then $\Delta\Delta C_T$ is the result of normalizing this ΔC_T value to a calibrator (e.g. time point 0) which is obtained by subtracting the C_T value of the endogenous control at the calibrating time point. So for the setup shown in **Table 11.1**, the $\Delta\Delta C_T$ value at 3 hours for phage gene D1 relative to time point zero, would be the C_T value for 3 hours minus the 16S value for 3 hours, minus (the C_T value for phage D1 time point zero – the C_T value for 16S time point zero). In summary, when this value is plotted as $2^{\Delta\Delta CT}$ then gives the amount of target transcript (phage D1 at 3 hours) normalized to an endogenous control (16S at 3 hours) and relative to a calibrator (phage D1 at 0 hours – 16S at 3 hours). In this approach, the standard deviations are calculated by taking the square root of the average calibrator value subtracted from the average sample value at any time point.

To measure the expression of phage and host genes during the infection cycle, I modified this protocol in two ways: (1) for a calibrator I used purified phage genomic DNA (this allowed the data to be presented in terms of transcript number) and (2) to this end, in order to maintain absolute values of the transcripts being assayed the standard deviation of the 16S mean C_T value was subtracted from the transcript assay C_T value (*see* **Note 4**).

An example template for Microsoft Excel of the modified analysis is shown in **Table 11.2**. The sample data shown are for the phage S-PM2 gene g47, in T4, gp47 is involved in recombination. Column 2 shows the average C_T value g47 at each time point (*see* **Note 5**). Column 3 shows the average 16S C_T value subtracted from the actual 16S value for each time point (*see* **Note 4**). The ΔC_T was therefore obtained by subtracting this (C_{T16S} – average C_{T16S}) from the gene 47 C_T value for each transcript (this is given in column 4). The calibrator (column 5) was the average C_T value of four phage genes (genomic template performed in triplicate) amplified from 10 ng of purified phage genomic DNA. Column 6 is the $\Delta\Delta C_T$ value obtained by subtracting column 5 from column 4. The next column *(7)* converts this value into the $2^{\Delta\Delta CT}$. The inverse of this value is given in column 8 which is the amount of transcript at the various time points relative to 16S rRNA and 10 ng of phage DNA. This value is then divided by 10 to account for the fact that 10 ng of phage DNA was used as template. Finally, the value in 9 is multiplied by 4,990,000 which is the number of copies of S-PM2 in 1 ng of DNA (*see* **Note 3**). Therefore, by plotting this value against the time points we now have a meaningful and useful way of looking at the number of copies of individual phage transcripts (*see* **Note 6**).

Table 11.2
An example of an Excel spread sheet designed to calculate the number of phage *psbA* (termed 'D1p' above) transcripts present at different infection times

1	2	3	4	5	6	7	8	9	10
Time	C_T (D1p)	Control (16S correction)	Corrected Δ C_T (2)–(3)	Calibrator C_{TAve}	$\Delta\Delta$ C_T (4)–(5)	$2^{(6)}$	Relative transcript copy 1/(7)	8/10	Absolute transcript copy × phage
1	20.058	−0.160	20.218	14.362	5.856	57.925	0.017263	0.0017263	8,614.587
3	17.993	−0.227	18.220	14.362	3.858	14.501	0.068959	0.0068959	34,410.61
6	19.798	0.429	19.369	14.362	5.007	32.158	0.031096	0.0031096	15,517.04
9	20.089	−0.278	20.367	14.362	6.005	64.227	0.015570	0.0015570	7,769.284

4 Notes

1. Unless absolutely necessary, duplexing reactions are generally not necessary.
2. Some master mixes contain UNG (Uracil-*N*-glycosylase which used in conjunction with dUTP prevents carry over contamination). UNG catalyzes the removal of uracil from uracil-containing DNA and thus ensures that any starting template is free from contamination of previously amplified products and generally is not necessary if good laboratory practice is maintained.
3. The number of phage genomes in the starting template of DNA used for real-time PCR can be calculated by dividing the nanograms of template used by the molecular mass of the phage and multiplying this value by Avogadro's number. To give an example for the scenario outlined in **Table 11.2**, if 1 ng of phage DNA is used as template and the phage has a genome size of 196,280, and the average molecular weight of a base pair of DNA is 615 Da, then the number of phage genomes = $[2 \times 10^{-9}/(196,280 \times 615)]\ (6 \times 10^{23}) = 4.99 \times 10^{6}$. NB: the reason this value is 2×10^{-9} is because DNA is double-stranded and thus twice the number of genomes are counted as one genome equivalent is amplified from each strand of DNA.
4. This is simply to use the phage genomic control as a calibrator and convert the value into absolute number of phage transcripts.
5. The C_T value at time point zero is 0 as no phages had been added at this point. This time point is not included in the table.
6. For this approach to work, all phage genes had to amplify equally efficiently at all C_T values. This proved to be the case and the standard deviation for all phage genes for a given template was negligible.

References

1. Heid CA, Stevens J, Livak KJ & PM., W. (1996) *Genome Research* **6**, 986–994.
2. Luke, K., Radek, A., Liu, X. P., Campbell, J., Uzan, M., Haselkorn, R. & Kogan, Y. (2002) *Virology* **299**, 182–191.
3. Poranen, M. M., Ravantti, J. J., Grahn, A. M., Gupta, R., Auvinen, P. & Bamford, D. H. (2006) *J. Virol.* **80**, 8081–8088.
4. Bustin, S. A. (2002) *Journal of Molecular Endocrinology* **29**, 23–39.
5. Dorak, M. T. (2006) *Real-time PCR* (Taylor and Francis.
6. Casas, V., Miyake, J., Balsley, H., Roark, J., Telles, S., Leeds, S., Zurita, I., Breitbart, M., Bartlett, D., Azam, F. & Rohwer, F. (2006) *Federation of European Microbiologica Societies: Microbiology Letters* **261**, 141–149.
7. Sandaa, R.-A. & Larsen, A. (2006) *Applied and Environmental Microbiology* **72**, 4610–4618.

8. Bookout, A. L. & Mangelsdorf, D. J. (2003) *Nuclear Receptor Signaling*, doi: 10.1621/nrs.01012.
9. Miller, E. S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T. & Rüger, W. (2003) *Microbiology and Molecular Biology Reviews* **67**, 86–156.
10. Clokie, M. R. J., Shan, J., Bailey, S., Jia, Y., Krisch, H. M., West, S. & Mann, N. H. (2006) *Environmental Microbiology* **8**, 827–835.
11. Mann, N. H., Clokie, M. R. J., Millard, A., Cook, A., Wilson, W. H., Wheatley, P. J., Letarov, A. & Krisch, H. M. (2005) *Journal of Bacteriology* **187**, 3188–3200.
12. Rebrikov, D. V. & Tofimov, D. Y. (2006) *Applied Biochemistry and Microbiology* **42**, 455–463.
13. Pfaffl, M. W. (2001) *Nucleic Acids Research* **29**, 2000–2007.
14. (2001) *User bulletin #2. ABI Prism 7700 Sequence Detection System*. (Applied Biosystems) http://docs.appliedbiosystems.com/pebiodocs/04303859.pdf.

Chapter 12

Oligonucleotide Microarrays for Bacteriophage Expression Studies

Andrew D. Millard and Bela Tiwari

Abstract

Gene expression microarrays offer the ability to monitor the expression of all phage genes over an infection cycle. However, there are relatively few reports to date of microarrays being used to investigate phage biology. This chapter aims to provide an overview of how to design and implement a microarray experiment to investigate phage biology.

Given the nature of microarrays being specific to an organism, each will provide a number of unique issues. In this chapter, we outline the basic theory behind microarrays and provide details on how to implement a microarray experiment from the design of oligonucleotide probes through to the hybridisation of microarrays. The matter of designing oligonucleotide probes will be discussed with regards to how probe length, secondary structure, free energy, probe orientation and amplification all have to be taken into account. As means of an example, the conditions used for the hybridisation of an array designed to be specific to the cyanophage S-PM2 is detailed.

Key words: Bacteriophage gene expression, bacteriophage microarrays, oligonucleotide probe design, hybridisation.

1 Introduction

Since their introduction *(1)*, microarrays have become popular and important tools in biological investigations. In general, microarrays are made from a solid surface, usually a membrane or glass slide, onto which stretches of DNA are bound in a regular array. Labelled nucleic acids isolated from the biological system under study are then hybridised to the microarray. Nucleic acids that bind to the sequences on the microarray are detected by measuring the label. In this manner, the sequence types represented in a sample can be identified and quantified. Many different types

Martha R. J. Clokie, Andrew M. Kropinski (eds.), *Bacteriophages: Methods and Protocols, Volume 2: Molecular and Applied Aspects, vol. 502*, © 2009 Humana Press, a part of Springer Science+Business Media
DOI 10.1007/978-1-60327-565-1_12 Springerprotocols.com

of arrays are possible and this chapter describes gene expression microarrays which are designed to measure RNA transcribed by genes represented on the array and are the topic of this chapter.

There are currently only a few published examples of bacteriophages being studied using expression microarrays *(2, 3)*. The expression of prophage genes has also been investigated using a microarray designed against the genome of *Salmonella enterica (4)*. The first microarray designed specifically to examine bacteriophage biology, rather than a bacterium containing a prophage, was for the archetypal myovirus T4 *(2)*. More recently, a study has been carried out to investigate two phages which infect the Gram-positive *Streptococcus thermophilus*. Both studies used arrays to determine the transcription of bacteriophage genes during infection of their hosts and were able to classify the majority of genes into early, delayed early, middle and late, based on their expression profiles. The popularity of microarray technology for studying bacteriophage biology is likely to grow; their small genomes and the dramatic change they can cause in host gene expression *(5)* make them ideal candidates for application of this technology.

The cost of microarrays is still high compared to techniques such as quantitative PCR and northern blots, but the prices of materials such as oligonucleotides for probes are decreasing (*see* **Note 1**) and the ability to get an expression profile for all the genes in a genome make this technique an attractive approach.

Microarrays are now a relatively mature technology and many books and articles have been written about their use (e.g. *see (6,7)*). For a comprehensive introduction to the use of microarray technology, we recommend the reader refers to books devoted to the subject. This chapter contains a basic introduction to designing a microarray and specific protocols for phage gene expression studies using microarrays. We describe the design and use of a microarray to study the cyanophage S-PM2, which infects marine *Synechococcus* as an example. We also discuss the potential, as well as limitations, of expression experiments using microarray technology.

1.1 Common Microarray Platforms

There are a number of microarray platforms, the two most common being high density oligonucleotide arrays and spotted arrays. The former uses a number of short, single stranded, short oligonucleotide probes bound to the array to represent a single gene, while the latter uses longer sequence tracts, with fewer representatives, often just one, per gene. The lengths of the probes (*see* **Note 2**) on these platforms differentiate them further. High density oligonucleotide arrays have short oligonucleotide probes, usually 25–30 bp long. Spotted arrays usually have one of two types of probe: 1) Double-stranded DNA probes. These may be generated using PCR amplicons of DNA (usually cDNA) from cloned inserts, or can be PCR products generated using

gene-specific primers with genomic DNA. These products vary in length, with the former producing probes that can range between approximately 200–2,000 bp and 2) Single-stranded, long oligonucleotide probes usually 50–70 bp in length.

Another key difference between array platforms is that only one sample is hybridised to each high density oligonucleotide array, while it is most common for two samples, each labelled with a different fluorescent dye, to be hybridised to each spotted array.

Issues to take into account when deciding what platform to use include the research question of interest, the organism being studied, and what resources, experimental and financial, are available. For most people researching phage biology, custom arrays as opposed to commercially mass produced arrays, will be the only choice. Although there is scope to design custom high density oligonucleotide arrays, this will be a realistic choice only for those who have the financial resources and plan to run large experiments using a particular chip design. For the majority of researchers, custom spotted arrays will be the platform of choice. There are a number of considerations when choosing between probe types for a spotted array. The main issues will be whether genome sequence for the phage of interest and its host are known and the cost and time implications of the available choices (*see* **Table 12.1**).

The phage studies mentioned in the introduction use microarrays with printed PCR products or PCR products in combination with short oligonucleotides as probes *(2, 3)*. Today, the benefits of using long oligonucleotide probes is such that for unsequenced organisms with small genomes, it may be more cost and time effective to invest in full sequencing instead of generating the materials necessary to make double-stranded probes from cloned inserts. The rest of this chapter considers issues and processes involved in generating and using microarrays with long oligonucleotide probes for expression experiments.

1.2 Designing Oligonucleotide Probes

An ideal oligonucleotide probe is specific to the gene (or gene family) of interest and is sensitive enough to detect low levels of a target sequence. Key considerations when designing oligonucleotides probes include probe sequence characteristics, such as length, melting temperature and potential for forming secondary structures, as well as characteristics of the probe in relation to its target sequence and the other, non-target sequences potentially present during an experiment. These include percentage sequence identity with non-targets, the position and length of matching sequence stretches in non-targets and binding free energy.

Choosing methods for finding good oligonucleotide probes, and deciding on parameters to give to those methods is challenging. If one very stringent criterion is applied, many suitably specific probes could be discarded, but applying a single relaxed

Table 12.1
Summary of double-stranded probes from cloned inserts versus long, single-stranded oligonucleotide or double-stranded PCR product probes from DNA

Factor	Double-stranded probes from cloned inserts	Long, single-stranded oligonucleotide or double-stranded PCR product probes from DNA
Basic requirements	Materials and expertise to create appropriate clone libraries and undertake all associated activities to generate probe material.	The sequence of the organism, and often also its host, needs to be available.
Speed to design and generate	If materials such as appropriate clone libraries already exist, a cDNA array should be quite quick to design and generate.	If genome sequence is available, probes or PCR primers for most genes can be quickly designed using appropriate software.
Convenience to design and generate	If libraries are not already available, this is a significant undertaking in time and resources. Quality control must be very strict, including planning for the storage and tracking of materials.	Usually long oligonucleotides or PCR primers will be designed by the researcher, or by a facility or company. Oligonucleotide probes or PCR primers are then generated to this design by a company. For PCR product probes, additional time and expertise is required to generate the PCR products for spotting.
Specificity	Without sequence information, the potential for cross-hybridisation of probes with non-targets is not known. Due to the length of these probes, it is usually assumed to be small.	Oligonucleotide probes can be designed to have particular specificity and sensitivity levels, and parameters such as desired length and melting temperature can be set during the design process. In addition, it is possible to design probes to detect gene families *(12, 20, 23)*, single nucleotide polymorphisms *(51, 52)* and splice variants *(12)*. Theoretically, PCR primers could also be generated to produce products with certain specificity and sensitivity levels, but even with software built to automate this task, there are quality issues such as ensuring that a product is clean and contains a single sequence type.
Future availability	Regenerating materials for additional print runs may require additional time and money.	Oligonucleotides and PCR primers can be regenerated easily. Generating and quality checking PCR products requires additional time.
		Probe and PCR primer information may already be available for some organisms. With the growth of probe databases such as that at the NCBI, it will become increasingly possible to download oligonucleotide probe sets for particular organisms or related species as well as to learn of experiments that used particular probes or probe sets.

(continued)

Table 12.1 (continued)

Factor	Double-stranded probes from cloned inserts	Long, single-stranded oligonucleotide or double-stranded PCR product probes from DNA
Cost	These probes may be the cheapest option if the array is to be used as part of a project where the required materials will be generated anyway. However, for projects starting out from scratch, this can be a lengthy and expensive process.	The price of oligonucleotide production is decreasing; for organisms with a fully sequenced genome, long oligonucleotide probes should prove a cost-effective option and is likely to be a more attractive solution for most researchers than using PCR products.
Quality control	There are many opportunities for contamination of samples and associated problems with annotation. *(53, 54)*. Quality control of all stages is vital.	Quality control is important but there is less scope for errors due to cross contamination of samples and errors in spotting locations for long oligonucleotide probes. PCR product probes require careful quality control and tracking.

criterion could predict many non-specific probes, which would be a big problem if these probes were used on a microarray.

Software for designing oligonucleotide probes employ a series of tests to detect acceptable stretches of sequence. Empirical results from He et al. (2005b) suggest that an appropriate combination of some measure of sequence identity between a probe and its target, the length of stretches of sequence identity with non-targets and a measure of free energy is sufficient to select specific oligonucleotide probes. **Table 12.2** is adapted from He et al. (2005b) and gives a summary of the minimum criteria they determined for designing gene-specific probes. Each of these criteria is discussed in further detail below along with additional considerations.

1.3 Probe Characteristics

1.3.1 Probe Length

There are a number of studies addressing the use of 50–70 bp long oligonucleotide probes *(8–14)*. Probe lengths of 70 bp and less are favored due mainly to issues associated with cost, specificity and the coupling efficiency during the synthesis of the oligonucleotides (*see* **Note 3**).

Evidence suggests that 70-mer probes give more reliable results than shorter probes. In a direct comparision of oligonucleotides of different lengths, 70-mer probes gave the most comparable results to cDNA probes and less target was necessary to get reliable results with 70-mers than 50-mers *(10)*. Signal intensities for 70-mer probes bound by target are greater than for 50-mer probes, with the difference particularly marked for genes expressed at a low level *(15)*.

Table 12.2
A summary of the essential probe design criteria for 50 and 70 bp oligonucleotide probes determined by He et al. (2005b)

Probe type and parameter	50-mer value	70-mer value
Maximum identity with non-target sequences	85%*	85%
Maximum stretch length of matching sequence with non-target sequences	15 bases	20 bases
Maximum binding energy with non-target sequences	−30 kcal/mol	−40 kcal/mol
Maximum number of self-binding oligonucleotides	8	8

*Tiquia et al. reported that 50-mers with greater than 75% similarity may show some cross-hybridisation.

1.3.2 Melting Temperature and Secondary Structure Potential

Ideally, all the probes on an array will have similar melting temperatures when bound to their targets, and their melting temperatures when bound with their most similar non-target sequences will be much lower. Some oligonucleotide design software will allow you to set a range of oligonucleotide lengths to make it easier to design probes with very similar melting temperatures. This is very useful, but it is important to find out from the company supplying your probes whether they will manufacture probes of various lengths or not.

When designing probes for bacterial genes, we found that the only way probes for bacterial genes could be designed to have similar melting temperatures to phage genes was to decrease the length of the oligonucleotide. Whilst many factors will effect probe melting temperature, one of these is the mol G+C content. The cyanophage S-PM2 has a mol G+C of 37.8 % which differs significantly from the 59.4 % of *Synechococcus (16)*. This is likely to be a problem with many phage host systems as significant differences in mol G + C content are common *(17)*.

Sensitivity may be affected if a probe can potentially form secondary structures. For probe design, it is the *potential* for secondary structure formation, rather than the most likely or stable structures that is the issue of interest. Potential for secondary structure formation in the target sequence may also affect sensitivity *(18)*. Predicted melting temperatures and probe secondary structure are discussed further in the 'Choosing Software' section of this chapter.

1.3.3 Sequence Identity

The specificity of an oligonucleotide probe for its target is determined by the number of mismatches between the probe and

target, the location and arrangement of those mismatches along the probe sequence *(9, 12)* and the similarity of the probe to any non-target sequences that may be present. Probes that cross-hybridise with sequences other than their target(s) can produce misleading, though highly reproducible results. Thus, probes should have little or no sequence similarity to non-target genes that may be expressed in the sample. Absolute levels of probe–target sequence identity necessary for signal to be observed depends on the stringency of the hybridisation conditions used. For example, increasing the temperature from 50 °C to 55 °C leads to the differentiation of sequences with higher levels of sequence specificity *(8, 19, 20)*.

In studying bacteriophages where gene-encoding RNA sequences do not have a polyA tail, the whole sample may be tagged with an oligo dA tail (**Section 1.3.5**). In this case, non-coding as well as coding RNA will be taken through the amplification and labelling process. Thus the potential presence of non-coding sequences from the organism of interest, as well as from other sources such as a host, needs to be considered during the oligonucleotide design process. For phage studies, where coding and non-coding sequences will be hybridised to an array and the presence of host sequences may be an issue, probe specificity should be checked by looking for (non-target) regions in both the phage and the host *genomes* with regions complementary to probe sequences.

A number of studies have investigated how similar non-target sequences need to be before significantly affecting microarray results. For 50-mer probes, sequences with greater than 75% similarity may show some cross-hybridisation *(8)*, while for longer probes such as 70-mers, the percentage similarity required for cross-hybridisation signal to become a problem is 85% or higher *(9, 19, 21)*. Additionally, long stretches (> 15 bp) of complementary nucleotides between a probe and a non-target sequence can lead to greater signal intensities than observed when a probe binds a different non-target sequence with the same overall percentage identity, but where the matching nucleotides are spread along the sequence. Accordingly, when multiple mismatches cannot be avoided during probe design, they should be spread across the length of the probe *(12, 19, 22)*.

Probes can also be designed to detect the presence of gene families *(12, 20, 23)*. Probes for this purpose should contain a long stretch of sequence common to all members of the group, while keeping stretches of identity with non-group members as short as possible. He et al. (2005b) provide some specific recommendations for this type of probe design.

1.3.4 Free Energy

Free energy can be used as a measure of oligonucleotide specificity in place of sequence identity. This measure takes into account

a number of factors such as matches, mismatches, the nature of adjacent nucleotides and interactions between the probe, targets and non-targets and so should be somewhat more sensitive than identity alone *(21)*. However, for sequence identities greater than 75%, minimal free energy is closely correlated to probe–target identity *(19)*. As it is recommended that sequence identity for gene-specific probes be greater than 85%, it may not be important whether one uses free binding energy measures or sequence identity measures when designing probes.

1.3.5 Amplification and Probe Orientation

Depending on the bacteriophage and host that are being studied it might be necessary to amplify RNA before it can be used in hybridisations. If amplification is required, then the amplification method should introduce no, or at least minimal bias. The strand of the target molecules in the final solution must be considered so that oligonucleotide probes for the microarray can be designed with the appropriate directionality to bind to the target molecules.

The two main approaches to increase the amount of target material are linear RNA amplification and logarithmic PCR-based amplification. The advantages of PCR-based methods are that they can amplify picogram quantities of sample and produce double-stranded products that are both more stable than the single-stranded products of the linear RNA amplification procedures and that they can be hybridised to probes in either orientation. However, PCR methods are sensitive to small changes in experimental conditions and may produce biased samples that do not represent the composition of the original sample *(24)*.

When working with most bacteriophages, it should be possible to extract the nanogram quantities of RNA that are required for a single round of linear amplification. This should yield enough RNA for most experimental designs, if not a second round of linear amplification can be applied. Many studies have shown that minimal bias is possible using a linear amplification method *(25–30)*, with a single round of linear amplification representing the target population well, while a second round can introduce small amounts of bias *(31)*, though depending on the experimental aims, this may be within acceptable levels *(30)*.

As in other stages of the microarray experimental process, the quality of the starting total RNA template is vital, as is the concentration of the promoter primer used to synthesise cDNA *(26)*. We refer the reader to a recent review of amplification techniques *(32)* and the possible implications for microarray studies *(24)*. The experimental steps taken need to be carefully considered in light of the questions being asked and the analysis and interpretation that is undertaken *(33)*.

The method of amplification will influence the design of oligonucleotide probes. The next choice is whether or not to convert RNA to the more stable cDNA. In our study, we used a

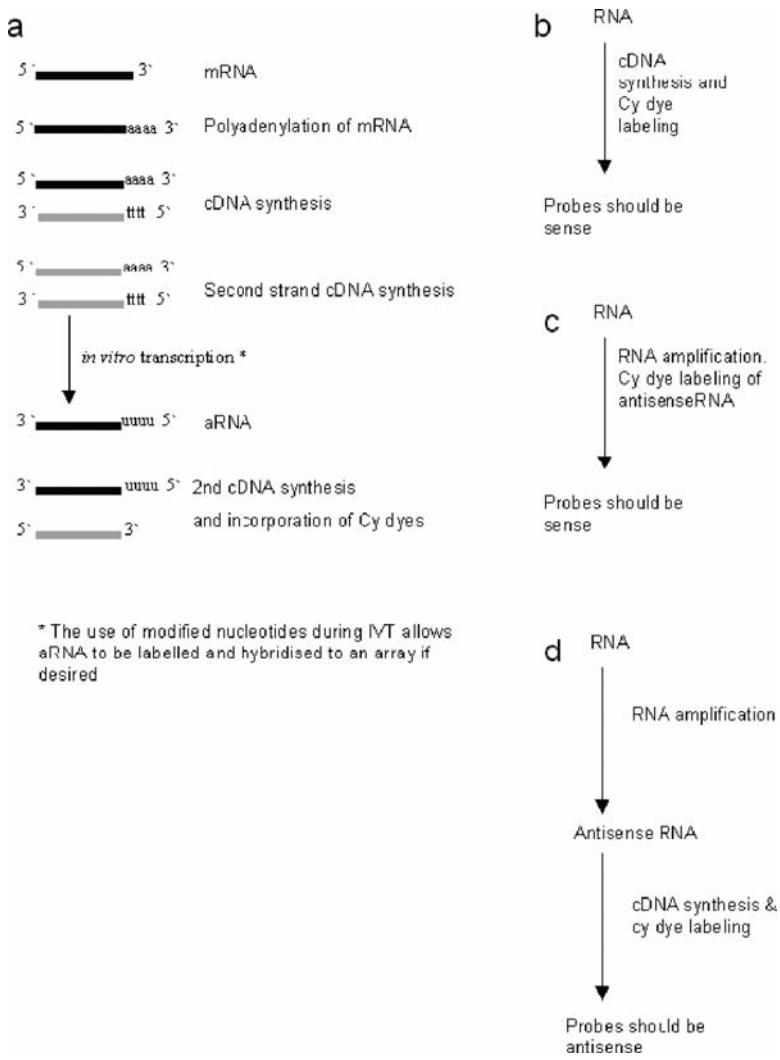


Fig. 12.1. Amplification of RNA directly affects the design of oligonucleotide probes. RNA was first polyadenylated which allows the Eberwine method *(56)* to be used to amplify RNA to produce antisense RNA, followed by synthesis of cDNA again (a). The choice of amplification and whether to hybridise cDNA or RNA to an array will affect the 'direction' the probes that have to be designed (b–d).

commercial kit for the linear amplification of RNA, the basic outline of which can be seen in **Fig. 12.1**. The choice of amplification method and what is hybridised (RNA or cDNA) to the array has to be determined before probes are designed. For our study, after the amplification of RNA, cDNA was synthesised which was then hybridised to the array. Therefore, the probes were designed to be antisense. If the resulting antisense RNA from amplification were hybridised to the array, the probes would have to be designed in the sense orientation. Therefore, it is essential to first determine how much RNA can be extracted from an infected culture

(*see* **Note 4**), if amplification is required and if cDNA or RNA is to be hybridised to the array.

1.3.6 Location of Probe Along Target Sequence

Probes can be designed to match any region of a target sequence. The amplification and labelling protocol used is key to the region to design probes against. Oligo dT priming creates a pool of cDNA biased towards sequence at the 3′ end of genes. This is especially true if amplifying from a small amount of original RNA material *(31)*. Random priming will produce sequence that can, at least theoretically, encompass the entire coding region from the points where primers bind to the 5′ end of each sequence. Thus in this case, more sequences will be represented by tracts closer to the 5′ end than the 3′ end. Accordingly, oligonucleotides should be designed closer to the 5′ end of the RNA for maximum sensitivity when using random priming *(34)*.

Recently, it has also been shown that the effect of the position of the fluorescent label relative to the probe–target duplex may have an effect on the signal intensity recorded for a given probe–target pair *(35)* although the effect is not large for long oligonucleotides—for the 50-mer probes used in that study, the signal intensity increased by a factor of less than 4 when the end-label of a target was close to the probe binding site rather than far away from it.

1.3.7 Replicate Spots

Printing replicate spots on an array is recommended as it gives scope for obtaining more precise estimates of signal intensities. In addition, using information from replicate spots helps distinguish signal due to target hybridisation from experimental artifacts such as faults with the surface of the slide.

The number of replicate probes printed on an array should be decided by balancing the amount of additional information, these spots provide with the cost of including them. To make analysis tractable, replicate spots should be printed with a constant distance between them across the array. If replicating entire print tip groups on the array, this will happen by default.

Many analysis programs average the log-intensities or log-ratios of replicate spots on an array and use these averages as an input to statistical analysis. However, more sophisticated methods that can increase the power of detecting differential expression have been developed for handling such data *(36,37)*.

1.3.8 Number of Probes per Gene

Multiple specific oligonucleotides representing a single gene can be spotted onto an array and many probe design programs allow the user to ask for multiple probes per gene to be designed. However, it is most common to spot a single probe per gene on custom spotted arrays. Hughes *et al.* reported that differential expression ratios obtained for a single 60-mer oligonucleotide probe were highly correlated to the average ratio for multiple 60-mer probes

per gene and with ratios from cDNA probes. Similar results have been reported for 50-mer probes although here the comparison was to short cDNA probes only (322–393 bp) *(12)*. Thus, depending on the experimental question, a single probe per gene may provide adequate and representative information about gene expression. If more than one spot per gene is included, it is important to consider how the data will be analysed to get representative information for a given gene.

1.3.9 Control Spots

Control probes are those for which target sequences will definitely be present, sometimes in known quantities, in the hybridisation mixture. Common controls are so-called 'housekeeping genes', which are assumed to be expressed at similar levels across all samples and exogenous nucleic acid, which may be spiked into the mixture at some point in the process of sample preparation. Control spots are useful for basic checks on print or hybridisation quality and even determining the orientation of the array. If a control target sequence has been taken through the RNA isolation process, it can facilitate evaluation of the quality of the RNA isolation and amplification processes.

Vitally for small arrays such as those used for studying phage expression, control spots can be used to normalise the data, which is a necessary step before statistical analysis. Common normalisation methods for spotted arrays, such as the Lowess method *(38)*, make the assumption that only a small percentage of genes are changing in their expression levels or that changes in expression are roughly symmetric between the two samples hybridised to a chip. These assumptions do not hold for arrays with only a small number of genes expressed at a constant level. Instead, the data can be normalised to spots where constant, preferably known, concentrations of target sequence are present in the hybridisation mixture. These targets of known concentration can also be used to determine the dynamic range and sensitivity of the system, which can be an important consideration in interpreting the data.

There are a number of commercial control probe sets available. These may be double-stranded PCR products or oligonucleotide probes. For studies on phage and other non-model organisms, the origin of the sequences provided in these sets should be investigated. Most claim to be species independent, that is, that they will not cross-react with the species being studied. However, this usually means that the controls have been screened against human, rat, mouse, yeast, and certain plant and bacteria species. The control sequences are not usually made available and companies may not even be willing to compare their control sequences with a particular genome to check for potential cross-hybridisation. However, certain companies may be willing to generate oligonucleotides meeting specifications laid out by a researcher. In this case, corresponding materials to spike into the

hybridisation mixture must also be generated. It is more difficult to use this route to normalise data as quality control of the concentrations of reagents added to the hybridisation mixture must be carefully controlled.

It is worth discussing what control sequences to use and how the data generated from this will be handled with the company and/or array facility you are working with.

1.4 Probe Design Software

There are many computer programs available to aid researchers in designing appropriate oligonucleotides for microarrays, many of which are completely free or free to academic users *(23, 34, 39–50)*. The aim of such programs is make the task of designing specific, sensitive oligonucleotides easier and faster. In general terms, they attempt to predict probe sequences that will bind well to target sequences (sensitive) and which will not cross-react with other sequences that might be found in the sample (specific). Sensitivity is usually predicted through a measure of melting temperature, percentage GC content and sometimes measures of free energy and prediction of potential secondary structures. Specificity is usually predicted by comparing a given oligonucleotide to sequences other than the target it is designed to hybridise to, and testing for percent identity, contiguous regions of identical sequence and low complexity regions. Although they have the same aims, the programs available differ in the methods they use, their speed and how user friendly they are. In addition, some probe design software address particular design issues such as finding probes specific to gene clusters as well as gene-specific probes *(23)*.

1.4.1 Choosing Software

To a person new to this area, and especially someone new to bioinformatics, just choosing software can seem a daunting task. Below we list a few things to consider when making a decision about which software to try.

1.4.2 Usability

The free oligonucleotide design programs available vary according to the platforms they can be run on (e.g. Windows, Mac OS, Linux), the quality of the documentation, and the type of user interface (command line, graphical, or web-based; *see* **Notes 5** and **6**). The interface that suits you depends on your computing experience and requirements. Biologists often want a good graphical interface with easy to understand outputs and reasonable speed.

1.4.3 Functionality

At first, it can appear that most oligonucleotide design applications are doing very similar things and that the user interface is the main differentiating factor. However, there are differences in their approaches to design that can influence how many and which oligonucleotides are chosen and how well these meet

experimental requirements. In addition, understanding, at least in general terms, the steps being carried out by the software can be important for troubleshooting and for interpreting array results.

A key example of different approaches and potentially different results due to them is associated with specificity checking. Much of the available software relies on Blast *(39)* for this step, but Blast was designed for searching large sequence databases at high speed; this is not the same task as specificity checking for oligonucleotide probes. Even when Blast parameters are tweaked to try and account for this, Blast can end up selecting sequences that are too similar to others in the sample as well as eliminating potentially good candidate oligonucleotides early in the process. Examples of recent, user-friendly software that have taken this into account and use custom sequence similarity searches are Picky *(40)* and Yoda *(41)*.

The methods used for predicting secondary structure potential also differ. Some use thermodynamic approaches while others merely check for complementary sequences that could lead to stem/loop structures. It is likely that the main difference worth considering for the purposes of oligonucleotide probe design is whether the program relies on external software for this process and how much time this step involves.

Melting temperature prediction for probe–target pairs is used to help design probes that will bind tightly to their target sequences. The aim in setting a particular melting temperature or range of temperatures that all probe–target pairs must have is to get probes that will behave consistently under the same hybridisation conditions. Thus, it is worth aiming for reasonably high predicted melting temperatures for a set of oligonucleotides (e.g. $> 63\,^{\circ}C$), and for reasonably consistent melting temperatures.

Most available software allows the user to set the melting temperature that a probe–target pair must attain to be considered, or a range it must fall within (*see* **Note 7**). There are variations on this theme where it is recognised that the absolute melting temperature of any given probe–target pair is not as important as the specificity of probes within a given experimental system. For example, the program Yoda calculates the average melting temperature of all possible oligonucleotides in a sequence set and uses this as a base around which a melting temperature range is considered. The Picky program takes a different angle, choosing oligonucleotides with at least a minimum difference between the melting temperature of a probe–target pair and the melting temperature of that probe with the most similar non-target sequence.

Good software also gives the user control over other key conditions such as which end of a sequence to preferentially design probes to, preferred oligonucletide sequence length and number of probes to design for each sequence. Other options to look for are the ability to automatically search for specificity against

sequences that are not part of the target set (e.g. host sequences), and to search against the reverse of sequences. The software packages mentioned earlier, Picky and Yoda, both allow the user to enter host sequences as well as target sequences for specificity checks, and both allow checks against the reverse of these sequences. In addition, both are fast, provide on-screen information about the progress of the job and produce tab-delimited text files of predicted oligonucleotides. Yoda's output is especially easy to use for iterative searches for oligonucleotides.

1.4.4 Practicality

Basic issues to consider include what format the input sequences need to be in, what type of results are saved, what format results are saved in and how fast the program runs. The documentation for each program should outline all of this information. Most programs accept sequences in FASTA format and at a minimum will return a tab delimited text file containing information about the oligonucleotides chosen.

The cost of the software should also be considered. The functionality, speed and usability of some of the freely available software are very good and should meet the majority of people's requirements. If a commercial solution is being considered, its usability and functionality should be compared with some of the free solutions before making a decision.

1.5 Practical Considerations

1.5.1 Working with a Facility

In general, we recommend that you work with a microarray facility to generate your arrays. Facilities usually provide advice about all aspects of array generation including oligonucleotide synthesis and other issues associated with the materials and methods they employ. They should be able to provide you with details of all the protocols and reagents they use to make your array. Some facilities may work with you to design your oligonucleotides, but others may assume that you will do this step yourself. When you are designing your first array, it may not be clear what questions to pose to the facility you are working with. Some suggestions are:

- Is the facility tied to a single supplier for oligonucleotides? Who is this and do they have any conditions that should be taken into consideration? For example, will the company supplying your probes manufacture probes of various lengths for you? Can you supply the facility with oligonucleotides from another source if you wish to?
- What are the minimum and maximum numbers of arrays that can be printed in a batch?
- How long does it take to get an array made from the start of design to the delivery of the printed slides?
- What quality control methods are carried out on a batch of arrays and in what format will these results be reported to you?

- What control spots will be included on the array and how do they recommend these be used during pre-analysis and analysis?
- What types of slides are used, what types of oligonucleotides (modified or unmodified) should be ordered? What types of blocking agents need to be included in the hybridisation solution?
- Will training for experimental techniques such as hybridisation, washing and imaging be provided?
- What support will the facility provide after the arrays are sent to you?

1.5.2 Understanding Spotted Arrays

This section gives a cursory overview of issues that should be considered before embarking on designing a microarray experiment using spotted arrays. We highly recommend that the reader refer to some of the books and articles available about the topics that follow to become familiar with the options available and their practical implications.

1.5.3 Signal intensity Versus Target Abundance

It is usual to hybridise two samples to an array, where each sample has been labelled with a different fluorescent dye. The log ratios of the signal intensities of the two colour channels are usually used during analysis. Comparisons of spot signal intensities, or their log ratios, give only qualitative indications of the amount of transcript in a sample. This is because there is much we do not understand about the non-linear dynamics of DNA and RNA hybridisation and the effect of the many sources of variation during hybridsation. Thus, we can use the results of spotted microarray experiments to determine if a given transcript is present, and if it is present in higher quantities in one sample than another, but we cannot infer its concentration or compare its abundance to other transcripts present in that, or any other sample.

1.5.4 Common Micro Array Designs

There are a number of standard microarray experimental designs *(42, 43)*. The design should be based on the experimental aims, material resources and statistical expertise available to the project. If you are new to this area or statistics, we highly recommend that statistical advice is sought during the design stage of the experiment.

Generally speaking, two types of comparisons can be made between samples: direct and indirect. Direct comparisons are possible when two samples to be compared are labelled with different dyes and hybridised to the same slide. Here, the log ratios of the signal intensities for any given spot are a measure of the abundance of the target transcript for that probe in one sample *relative to* that target in the other sample. It is most common to use log ratios when analysing two-colour spotted arrays. Recently, examples of analysis using only one colour channel have been

published. These are mentioned later in the text, but for clarity, we refer only to log ratios.

For indirect comparisons, a single log ratio is generally not informative; it is the log ratios from different slides that provide information about relative target abundance in the samples. For example, the 'reference' design, commonly used in microarray studies, involves indirect comparisons: a single reference sample (*see* **Note 8**) is labelled with one dye and hybridised to all the slides in an experiment, with each experimental sample labelled with the other dye and hybridised to one slide per replicate. The resulting ratios of sample to reference are then compared to give indications of the relative amounts of particular transcripts in each sample. Levels of variation are higher when using indirect comparisons relative to direct comparisons *(44)*, but the practical benefits of using indirect comparisons can outweigh this issue. Common microarray experimental designs are described briefly in **Table 12.3**.

1.5.5 Confounding Effects (Lurking Variables)

The usual aim of an experiment is to measure some biologically meaningful effect. It is important that no uninteresting factors, such as which person prepared the RNA samples, or which day the experiment was run, contribute to conclusions about the biological system. For example, if RNA from all treated samples is collected by Bob, and RNA from all control samples are collected by Mary, how can we tell which effects measured in the experiment are due to differences between the treatment and the control, and which are due to differences between how Bob and Mary handled the samples? This type of problem is referred to as confounding and it is important that all such possible factors be considered and removed through design (blocking) or randomisation before undertaking the experiment.

An easily overlooked source of possible confounding is the array itself. Evidence suggests that arrays printed later in a print run may contain less probe than those printed earlier in the process and that this can lead to lower signal intensities and greater variation in signal intensity ratios *(34)*. If all samples of one type are hybridised to arrays printed early in the print run, and all samples of another type are hybridised to arrays printed later in the run, the order of printing may be contributing to identified effects. Similarly, it is important to use arrays from the same batch for an experiment whenever possible.

Good recording of all experimental details is important, as many issues that become apparent during analysis can be investigated later. For example, an investigator reported that their array results seemed different for one part of an experiment they were running. By talking with all the people involved, they managed to determine that the RNA isolation protocol used had been slightly different for the samples affected. If such detail is recorded for

Table 12.3
There are a number of common microarray experimental designs to choose from. Their benefits and drawbacks, including experimental aims and cost implications, should be considered before deciding which approach to take

Design Type	Type of Comparison	Experiment sizes	Variation	Analysis issues	Other comments
Direct ratios	Direct	Good for small numbers of treatments (e.g. <4)	Smallest amount of variation in resulting log ratios.	For more than two samples, this can be treated as a loop design, but some software will not handle the analysis. Robust if enough replication.	Ratios are not directly comparable with experimental results generated during separate experiments.
Reference design	Indirect	Used for experiments with a greater number of treatments or where results across experiments will be directly compared.	Larger amount of variation than direct comparison.	Easy to analyse.	Robust but inefficient design; the loss of one chip is not usually a problem for analysis, but extra resources are being dedicated to hybridising reference materials. Need to generate a standard reference that can be used across the experiment (and later experiments if desired).

(continued)

Table 12.3 (continued)

Design Type	Type of Comparison	Experiment sizes	Variation	Analysis issues	Other comments
Loop design	Direct and indirect	Good for small to medium numbers of treatments.	Lower variation for the samples compared directly.	Most important comparisons can be done with direct comparisons, while less important comparisons are done through indirect comparisons.	Often not very robust. The loss of a single slide can cause problems although there are modifications to the loop design to try and address this. A good level of statistical understanding and an ability to use statistical software is necessary.
One-colour design	Indirect	Could be applied to any size of experiment.	Larger amount of variation than direct comparison and very dependent on the high standards and consistency of the protocols and materials throughout the experiment.	Easy to analyse.	Not commonly used with spotted arrays as it does not account for the many sources of variation affecting any single array. However, some authors report that results are comparable to other indirect comparisons in terms of variation and reliability *(44, 55)*

each sample used in an experiment, it is relatively straight forward to determine the cause of anomalies like this and to decide how to move forward with analysing the data (*see* **Note 9**).

1.5.6 Replication and Pooling

Replicates are vital to microarray studies. A replicate is a repetition of the same trial. There are two main types of replicates: biological and technical. If the exact same samples are hybridised to a slide, it is a technical replicate. Such a replicate does not give information about what is happening in the population of samples, only a more precise measure of what has happened in that one sample. If samples from the same treatment groups, but not the same individuals, are hybridised to a slide, it is a biological replicate. Biological replicates allow estimates to be determined for the expression of genes in a population; they do not necessarily give a precise measure for what is occurring in a given individual. For microarray studies, technical replicate hybridisations should be done with material separated as early in the experimental process as possible. For example, if replicates are generated by taking labelled sample, splitting it into two batches and hybridising to two slides, a very precise measurement for each gene in that sample is obtained, but no information is obtained about the range of values possible due to technical variation such as pipetting during the RNA isolation or labelling.

For lab-cultured organisms, the question arises: what is a biological replicate? If a phage population has been cultured for a long time, all samples used in a microarray study may be thought of as 'biological' or 'technical' replicates, depending on the length of time after sub-culturing, the organism's biology and the terminology common in that area of biology. The important point is to keep in mind the biology of the organism and how it has been treated when analysing and interpreting results rather than assuming that the terms 'technical replicate' and 'biological replicate' are able to adequately describe the samples in your experimental system.

Another issue when working with populations of organism that are treated as samples is that every microarray hybridisation contains material from a pooled sample. Thus array results give no indication of what an individual phage is expressing, rather they indicate what the population of phage from which RNA was isolated was expressing. While this seems obvious when written down, it is easy to be blinded by the volume of data from an array experiment and should be borne in mind when interpreting and reporting results.

For statistical studies, the number of biological replicates required can be calculated using the minimum effect size to be detected, the tolerable error rates (false positives and false negatives) and a measure of the variation of the effect being measured in a power analysis. It is not easy to determine the required

number of replicates in a microarray experiment this way. A key problem is that many genes are being measured simultaneously, and each might have its own level of variation in the population. In other words, the number of biological replicates required to reliably detect an effect might be quite small for genes with low inter-sample variability, but very high for genes with high inter-sample variability (*see* **Note 10**). For custom arrays, the cost of carrying out preliminary experiments to determine gene-level variation is often prohibitive and the reality is that many researchers calculate the number of replicates they can afford financially rather than on scientific or statistical principles. This is acceptable within reason if the implications of design decisions for the experimental aims are taken into account, and assuming that the microarray study is essentially being used to screen for potentially interesting genes. That is, assuming that genes identified as interesting through array studies will be further investigated using other experimental methods. For lab-cultured phage studies, biological variation will be quite low, so the required number of replicates will be lower than if more heterogeneous population was being studied.

2 Materials

2.1 RNA Extraction

Materials and methods for the extraction of RNA from phage-infected cultures have been previously described in detail elsewhere in this book in **Chapter 34**.

2.2 RNA Amplification

1. MessageAmp™ II-Bacteria RNA Amplification Kit and buffers (Ambion).
2. Ethanol ACS Grade (Sigma Aldrich).
3. RNase & DNase free water (Fisher).

2.3 cDNA Production

1. Superscript III (Invitrogen).
2. 5× forward stand buffer (Invitrogen).
3. aa-dUTP (Sigma) 100 mM solution.
4. A 5 µl of 100 mM dATP, dGTP, dCTP and 2.5 µl of aa-dUTP and dTTP were combined with 30 µl of H_2O to make a 20× stock solution of aa-dUTP/dNTPs, this was stored at −20 °C (*see* **Note 11**).
5. Random Hexamers (TAGN).
6. 80% Ethanol.
7. Nucleospin Extract II columns and buffers (Abgene).
8. A 10 N NaOH: 8g NaOH made up to a total volume of 20 ml with H_2O.

9. A 3 M sodium acetate: 81.6 g of sodium acetate dissolved in water with final volume of 20 ml. Stored at room temperature.
10. RT stop solution: 1 ml of stop solution contains 900 µl 0.5 M EDTA and 100 µl of 10 N NaOH. Stored at room temperature.
11. RT neutralisation solution: 1 ml of neutralisation contains 750 µl of 1 M HEPES (pH 7.4) and 250 µl of 3 M NaOAc. Stored at room temperature.

2.4 Cy Dye Coupling

1. Cy 3 and Cy 5 dyes (Amersham) were resuspended in 74 µl of DMSO and split into aliquots of 4 µl. Aliquots were used immediately or desiccated and stored at −20 °C protected from the light (*see* **Note 12**).
2. Stock solution of 4 M hydroxyammonium chloride was made with 0.27 g in 1 ml of H_2O, aliquoted into 50 µl volumes and stored at −20 °C.
3. A 0.5 M solution of sodium bicarbonate: 0.42 g in 10 ml H_2O. A 25 µl aliquots were made and stored at −20 °C.

2.5 Pre-washing and Blocking

1. 0.1% (v/v) solution of Triton X-100.
2. A 1 M HCl: 100 ml of 32 % HCl, 900 ml of H_2O.
3. A 1 M stock solution of KCl: 74.55 g dissolved in H_2O and made up to 1 l.
4. Blocking solution 50 ml: 38.43 ml of 4 × Block E (Nexterion), 11.53 ml of H_2O and 30 µl of 1 M HCl.
5. Eppendorf 5804 Centrifuge with rotor A-2-DWP (*see* **Note 13**).
6. Stainless Steel 25 glass slide rack.
7. Glass staining dish.

2.6 Hybridisation

1. Hybridisation chamber (Genetix).
2. Nexterion E Hybridisation buffer.
3. Hybridisation oven.
4. Lifter slips (VWR).

2.7 Post-Hybridisation Washes

1. A 10% (w/v) SDS stock solution: 50 g SDS made up to 500 ml with H_2O. 20 × SSC: dissolve 175.3 g NaCl and 88.2 g sodium citrate in 800 ml H_2O and adjust to pH 7.6, make final volume to 1 l and store at room temperature.
2. A 2 × SSC and 0.2% SDS solution: 100 ml 20 × SSC, 20 ml 10% SDS and 880 ml H_2O. Solution is made fresh on the day of use.
3. 2 × SSC: 100 ml 20 × SSC and 900 ml H_2O. Solution is made fresh on the day of use.
4. 0.2 × SSC: 10 ml 20 × SSC and 990 ml H_2O. Solution is made fresh on the day of use.

3 Methods

Before a microarray experiment can be carried out, a large amount of time needs to be spent on the preparation of high titre phage stocks. We used plaque assays to produce high titre lysates followed by caesium chloride gradients to purify the phage prior to use in the experiment (methods for these techniques have been described elsewhere in Volume 2 Chapters 13 and 14). This high titre stock was then used to inoculate a 1 L culture of *Synechococcus* from which samples were removed every hour for 10 hours.

3.1 Oligonucleotide Probes

The probes used on the S-PM2 microarray were synthesised by Operon with a 5′ amino modification for the attachment to the array. They were supplied in a 384-well plate (*see* **Note 14**). Oligos varied in length from 60 to 70 bp long. Oligos were designed in the antisense direction as cDNA was to be hybridised to the array and an RNA amplification step was included.

Oligos were designed taking into account the factors described in **Section 2.1**. At times it was not possible to design a 'good' probe, and a trade off was made between the data that might be gained from a 'poor' probe and having no data at all for a particular gene. As phage have the ability to pick up copies of genes from their host, problems can arise designing probes specific to particular phage mRNAs. For example, in the S-PM2:*Synechococcus* phage:host system, we were unable to design probes specific to a set of phage genes that are copies of host genes involved in photosynthesis *(45)*. In these cases, we designed probes that detect the presence of a host or phage RNA species and will keep this observation in mind when interpreting the data. Such differences will have to be teased apart using RT PCR (**Volume 2 Chapters 11**).

3.2 RNA Extraction

RNA was extracted using the protocol previously described in **Chapter 34** (*see* **Note 7**). RNA was checked for DNA contamination using multiple sets of PCR primers specific to both host and phage genes. RNA quantity was determined using a nanodrop spectrophotometer and quality was determined using a bioanalyser. (Details can be found in **Chapter 34**.) RNA with an A_{260}/A_{280} ration of 1.9–2.1 was used for amplification. Once RNA was extracted, it was aliquoted, snap frozen and stored at −80 °C until it was used in the amplification process (*see* **Note 15**).

3.3 Printing of Arrays

There are a number of different chemistries that allow the immobilisation of DNA on the slide surface for printing arrays (*see* **Note 16**). We used Nexterion E slides, which are epoxy coated.

The arrays were printed at a specialist facility. In most cases, arrays will not be printed 'in house' and a specialised facility will be used, where expert guidance can be gained on the choice of slide and method of printing.

3.4 RNA Amplification

An important factor in experimental design is whether enough RNA for a hybridisation reaction can be extracted from a sample without amplification. With the phage–host: system of S-PM2 and *Synechococcus*, it was possible to extract enough RNA for a single hybridisation. However, it was not possible to simply increase the volume of infected culture used to extract RNA from and the time taken to prepare high titre stocks of S-PM2 is prohibitive. Therefore, amplification of RNA was needed to follow the expression of phage genes every hour, for 10 h.

Before amplification all RNA was diluted (or concentrated) to a concentration of 125 ng/μl. MessageAmp™ II-Bacteria Kit from Ambion was used for amplification; the protocol described below follows that in the manual provided by Ambion with a few minor modifications.

1. For each sample, add 4 μl of sample RNA and 1 μl of spike RNA to a 0.2 ml PCR tube (*see* **Note 17**).
2. Incubate at 70 °C for 10 min, remove and cool on ice for 10 min.
3. While the RNA is incubating at 70 °C make a master mix for the polyadenylation step, for each sample 1.5 μl of nuclease water, 1 μl of 10× Poly(A) Tailing Buffer, 1 μl of RNAse inhibitor, 0.5 μl of poly(A) tailing ATP and 1 μl of poly A polymerase enzyme is required.
4. Add 5 μl of master mix to each tube, vortex briefly and then centrifuge to ensure the mix is at the bottom of the tube.
5. Incubate at 37 °C for 15 min in a PCR machine, then cool to 4 °C and transfer to ice.
6. Make a master mix for the first strand cDNA step. For each sample, 3 μl of water, 1 μl T7 Oligo (dT) VN, 1 μl of 10× first strand buffer, 4 μl of dNTP mix and 1 μl of Array Script for each reaction is required. Add 10 μl of this master mix to the PCR tubes and incubate for 2 h at 42 °C and then cool to 4 °C on ice.
7. Pre-heat an aliquot of water to 50 °C, with 20 μl needed for each amplification that has been performed.
8. Make a master mix for second strand cDNA synthesis. This requires 63 μl of water, 10 μl 10× second strand buffer, 4 μl of dNTP mix, 2 μl of DNA polymerase and 1 μl of RNase H for each reaction. Vortex, centrifuge briefly and add 80 μl to each tube and incubate at 16 °C for 2 h.

9. Immediately add the cDNA to 250 µl of cDNA binding buffer in a 2 ml Eppendorf and mix. Then pipette this mixture onto the membrane of a cDNA purification column.
10. Centrifuge at $10,000 \times g$ for 90 s, and discard the flow through.
11. Apply 500 µl of wash buffer to the cDNA column, centrifuge again at $10,000 \times g$ for 90 s. Discard the flow through and centrifuge for a further 2 min.
12. Transfer the cDNA column to a fresh cDNA elution tube and add 20 µl of pre-heated water to the column. Leave at room temperature for 5 min, before centrifuging at $10,000 \times g$ for 2 min to elute the double-stranded cDNA. Store the ds cDNA on ice.
13. Make a final master mix for the synthesis of aRNA by *in vitro* transcription (IVT). For each amplification 4 µl of T7 CTP, T7 GTP, T7 UTP, T7 ATP, 10× T7 reaction buffer and 4 µl of enzyme mix is required. Add 24 µl of this master mix to the eluted dscDNA (*see* **Note 12**).
14. Incubate at 37 °C for 14 h for the IVT and then transfer to ice. Bring the volume up to 100 µl with water (*see* **Note 18**).
15. Incubate an aliquot of nuclease free water at 55 °C, with 150 µl needed per reaction.
16. Add 350 µl of aRNA binding buffer to each sample and vortex gently, add 250 µl of 100% ethanol and use a pipette to mix the solution three times. Immediately pipette the above mixture onto the centre of an aRNA column and centrifuge for 2 min at $10,000 \times g$. Transfer the column to fresh elution.
17. After discarding the flow through, add 650 µl of wash buffer to the column, and centrifuge for 1 min.
18. Discard the flow through and centrifuge for a further 2 min, and transfer the column to a fresh collection tube.
19. Add 75 µl of pre-heated nuclease free water (55 °C) to the centre of the column, leave for 5 min and then centrifuge for 2 min. Add a further 75 µl of water and centrifuge for a further 2 min.
20. Discard the cartridge and keep the 150 µl of amplified RNA. RNA was quantified by the use of a nano-drop. The RNA was frozen at −80 °C or used immediately.

3.5 cDNA Production

cDNA is modified with an amino ally base modification that binds to the monofunctional NHS-ester Cy dyes to allow visualisation of the hybridisation of cDNA to the probes (*see* **Note 19**). If RNA is not amplified, the synthesis of cDNA is still necessary. However, the amount of RNA required for the cDNA reaction will be larger if unamplified RNA is used as it will contain a mixture of the mRNAs of interest and other forms of RNA.

For the synthesis of cDNA from aRNA, reactions were carried out in 200 µl PCR tubes. The resulting cDNA was

purified by the use of columns and buffers from the Nucleospin Extract II kit (Abgene) following the procedure developed by the manufacturer.

1. Assemble 2 μl of random hexamers, 2 μl dNTP mix, 1,000 ng aRNA and water up to a total volume of 28 μl in a 200 μl PCR tube. Incubate at 65 °C for 10 min, and then cool on ice for 2 min.
2. Add 8 μl of 5× forward strand buffer and 2 μl of 0.1 M DTT and incubate at 25 °C for 2 min. Add 2 μl of superscript III RT enzyme, incubate at 25 °C for a further 10 min. Then incubate at 42 °C for 50 min.
3. Add 10 μl of stop solution. Incuabte at 65 °C for 15 min.
4. Add 30 μl of neutralising solution and mix gently and transfer to a fresh 1.5 ml Eppendorf.
5. Add 400 μl of NT solution and mix by pipetting up and down, before transferring onto a Nucleospin Extract column.
6. Centrifuge the column at 13,000 rpm for 1 min at room temperature, and discard the flow through.
7. Add 800 μl of 80% ethanol to the column, and centrifuge for 1 min. Discard the flow through (*see* **Note 20**).
8. Add 200 μl of 80% ethanol and centrifuge at 13,000 rpm for 5 min. Discard the flow through and transfer the column to a new 1.5 ml Eppendorf.
9. Add 40 μl of ready warmed H_2O to the column, centrifuge at 13,000 rpm for 1 min. Add a further 40 μl of H_2O and repeat the centrifugation.
10. Split the cDNA equally between two 1.5 ml Eppendorfs, and dry the cDNA in a vacuum centrifuge.

3.6 Cy dye Coupling

Indirect labelling is used to incorporate Cy dyes. The cDNA is then purified before hybridisation using reagents from a Nucleospin II extract kit to remove any un-incorporated dye (*see* **Note 21**). The method for incorporation is outlined below.

1. Add 5 μl of 0.1 M sodium bicarbonate to each sample of cDNA to resuspend the cDNA pellet.
2. Add 5 μl of Cy 3 to the control and 5 μl of Cy 5 to the experimental sample, and mix by pipetting. Incubate at room temperature for 1 hr (see **Note 22**).
3. Add 6 μl of 4 M hydroxylamine to each sample and mix by pipetting. Incubate for 15 min at room temperature (*see* **Note 23**).
4. Add 70 μl of H_2O to the experimental sample and 400 μl of NT buffer to control sample. Ensure both samples are mixed and combine in single tube before adding to a Nucleopore Extract column.
5. Centrifuge the column for 1 min and discard the flow through.

6. Add 600 µl of NT3 buffer to the column and centrifuge for 1 min and discard the supernatant.
7. Add a further 200 µl of NT3 and centrifuge for 5 min, discard the supernatant.
8. Place the column in a fresh 1.5 ml Eppendorf tube. Add 50 µl of NE buffer and elute the Cy labelled sample by centrifugation for 1 min.
9. Concentrate this volume to *c.* 10 µl by the use of a vacuum centrifuge.

3.7 Pre-washing and Blocking

Slides are first washed and then blocked before hybridisation to ensure unbound probes are removed and there is no non-specific binding. Slides are washed in a glass bowl using a metallic slide holder and are handled by the edges to avoid contact with the oligo probes. Slides are not allowed to dry out between washes. All washes are carried out on a variable speed rocker to ensure constant movement of the solutions around the slides. The following washing procedure is for Nexterion E slides. For other types of slides, alternative washing procedures may have to be followed. These details should be provided by the facility that prints the arrays.

1. Fill a staining dish with 0.1% Triton X-100 at room temperature. Place the slides in the slide rack and immerse in the triton solution for 1 min.
2. Fill a fresh staining dish with 1 mM HCl. Transfer the slides into a fresh staining rack and immerse in the 1 mM HCl for 2 min.
3. Repeat this procedure with fresh 1 mM HCl for a further 2 min.
4. Wash the slides in 100 mM KCl for 10 min, ensuring the slides are transferred into a clean slide rack each time.
5. Wash the slides in H_2O for 1 min.
6. Transfer the slides to a plastic slide holder and fill with 1× Nexterion E blocking solution. Incubate at 50 °C for 15 min in a hybridisation oven.
7. Transfer back to a metallic slide holder and wash in H_2O for 1 min.
8. Place the slider holder in a lid of a 96-well plate and place in centrifuge. Ensure the centrifuge is balanced and centrifuge at 1,000 rpm for 1 min. Pour off any water that has collected in the lid and invert the slides. Centrifuge for a further 4 min.

3.8 Hybridisation

The hybridisation conditions detailed below are specific to Nexterion E slides printed with 70-mer oligonucleotide probes. For other types of slides, alternative washing temperature and washing solutions may be required. Hybridisations are carried out in a hybridisation chamber.

1. Prepare the hybridisation chamber at least 30 min before use. Fill the reservoir in the chamber with H_2O and seal the chamber. Incubate at 65 °C for at least 30 min.
2. Resuspend the 10 μl volume of Cy labelled cDNA in 90 μl of Nexterion E hybridisation buffer, mix by pipetting up and down. It may be necessary to warm the hybridisation buffer to ensure that the SDS is resuspended in solution.
3. Place in a hot block at 99 °C for 2 min.
4. Centrifuge the sample for 1 min.
5. Place a lifter slip over the oligo probes on the array, with the raised edges of the lifter slip face down. This forms a small reservoir underneath the slip. Pipette the hybridisation solution along the edge of one end of the lifter slip; the solution will be drawn under the lifter slip. Ensure that the solution is drawn all the way along the lifter slide.
6. Place the slides in the pre-warmed hybridisation chamber. Ensure a tight seal is formed and incubate the chamber at 65 °C for 14 h.

3.9 Post-Hybridisation Washes

Washes were performed in staining dishes with slide racks on a variable speed rocker. The dishes were covered in tin foil to minimise the amount of light reaching the slides.

1. Fill a staining dish with a 2× SSC, 0.2× SDS solution. Remove the slides from the hybridisation chamber. Place the array, lifter slide down in a 2× SSC, 0.2× SDS solution and gently agitate to remove the lifter slip. Then place in slide holder and immerse in the same solution for 15 min. Ensure that all of the slides are completely covered by the wash solution.
2. Transfer to a 2× SSC solution, and wash for 15 min.
3. Transfer to a 0.2× SSC solution, and wash for 10 min. Transfer to fresh a 0.2× SSC solution, and wash for a further 10 min.
4. Dry the slides in a centrifuge as previously described in **Section 3.7**. Place the slides in a slide holder and protect from the light.
5. Scan the slides as soon as possible, if they cannot be scanned immediately then store in the fridge at 4 °C protected from the light.

3.10 Testing Arrays

It is necessary to test an array before embarking on a full-scale experiment. The first test of an array should be to determine if there is any hybridisation of target to the array. The specificity of the probes can also be tested. It is possible to determine if host RNA will hybridise to 'phage-specific' probes by hybridising RNA from an uninfected host to slides; there should be minimal signal from good phage-specific probes. The identity of any probes that appear to be non-specific should be recorded and this information

used when analysing and interpreting experimental results. For spiked RNA controls, there will need to be some optimisation of the amount of spike added to cDNA synthesis or the amplification step. For commercial spikes, guidance is provided by the manufacturer.

4 Notes

1. Before embarking on an array experiment, the full cost of experiment should be calculated. The largest cost initially incurred will be the oligonucleotide synthesis. However, even the smallest scale synthesis will provide enough for the printing of hundreds of arrays. To aid in comparing the costs of different stages of an array experiment, consider the following: if the cost of oligonucleotides is taken as 100p, the price for printing of an array (including the slide) is *c*. 0.625p, the cost of RNA amplification is *c*. 1p per array, for Cy dye-labelling the cost is *c*. 0.4p array, and for cleanup and hybridisation is *c*. 0.25p. This means that if 50 arrays are used in an experiment, the cost of amplification and labelling soon exceed the cost of the original purchase of oligos.
2. The term 'probe' when used with microarrays refers to the nucleic acid bound to the microarray surface. 'Target' refers to the nucleic acid in solution, the presence and relative abundance of which is being measured during the experiment.
3. Longer oligonucleotides cost more to produce, but if designed well can have greater specificity for a target sequence than a shorter probe. However, as the oligonucleotide gets longer, the effects of coupling efficiency become more marked. For example, if the oligonucleotide synthesis coupling efficiency is 99%, a theoretical maximum of only 60.5% of 50-mer oligonucleotides will be full length. For 70-mers, this decreases to just under 50%.
4. As a guide it has been estimated between 20 and 70 μg of total RNA will be needed for each array *(46)*. It was found that the yield of RNA that could be extracted from *Synechococcus* cells infected with S-PM2 increased dramatically as the infection cycle progressed, this should be taken into account if trying to determine if amplification of RNA is necessary.
5. Some of the benefits and drawbacks of common software interfaces are described in **Table 12.4**.
6. If there are privacy or intellectual property issues associated with your sequences, using remote resources is not desirable. When accessing web-based or client/server oligonucleotide software where specificity is determined using Blast, a database of your organism's sequences (or

Table 12.4
Advantages and disadvantages of common software interfaces

Interface	Benefits	Drawbacks
Command Line	Easy to run if instructions are read Useful for incorporating into automated tasks or pipelines Usually secure	Not intuitive Not encouraging to new users Speed depends on local computing resources
Graphical (stand-alone)	Usually intuitive to run Biologists are usually comfortable using graphical tools Usually secure	Not always easily automated or incorporated into pipelines
Web-based (or graphical client software)	Easy to run Biologists are usually comfortable using web-based tools Remote resources used for computing	Not usually automatable. Often need to create an account. Data is sent to a remote machine (see **Note 5**) It can take a long time to receive results. If genomes must be available on the remote machine, non-model organisms may not be catered for (though the hosts of the service may add yours if you ask).

any other sequences probes should not cross-hybridise with) needs to be made available on that server. People running server-based software may be willing to add your database to their list if you request them to but, again, for any sequences that cannot be publicly released, such services may not be suitable.

7. The actual hybridisation conditions used in the experiment may not reflect the predicted melting temperatures used when designing probes. Most software predicts melting temperatures using the nearest neighbour method with parameters described by Santalucia *(47)*. This method assumes that the two nucleotide strands are in solution, with particular concentrations and conditions—assumptions that are not met in a microarray experiment.
8. Reference samples have commonly been made up of aliquots of the samples of interest pooled together. As some of the reference pool is hybridised to every slide and any transcript present in the experimental samples will be represented in the reference, signal intensity ratios should be calculable for each

transcript present at a detectable level. If data from experiments carried out at different times needs to be integrated or compared, then the same reference pool should be employed in the different experiments. This is not always feasible when using pooled aliquots of experimental material and, as the signal intensities for the reference samples are not in themselves informative, other types of reference material have been suggested such as a universal oligonucleotide *(44, 48)* and using genomic DNA *(49)*.

9. Electronic recording of details such as treatments and protocols will make it much easier to troubleshoot, especially if details are recorded in a relational database. We also recommend setting up a good naming scheme that makes obvious the providence of the material at each stage of the experiment. There are a number of databases available designed to handle microarray annotations and data. Such databases can also make annotation to MIAME standards and submission of data to public repositories such as ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) easier.
10. A positive step towards dealing with this challenge is a project using data available from the GEO database to help researchers determine power estimates or sample sizes *(50)*. This resource may aid those designing experiments involving organisms and genes used in published experiments. Hopefully, as more phage microarray experiments are published, tools such as this may be of more use to this community.
11. The water used in enzymatic reactions, in the dilution of RNA, for resuspension of primers and in the dilution of hybridisation buffer was nuclease free. All other is milliQ grade, that is it has a resistance of 18.2Ω.
12. The storage of Cy dyes in solution can lead to reduced coupling efficiency as water causes the hydrolysis of the Cy dye esters. As DSMO is hygroscopic, this is a distinct possibility.
13. Other centrifuges and rotors can be used, as long as the rotor is capable of centrifugation of 96-well plates.
14. When choosing a manufacturer of oligonucleotides, a number of factors need to be taken into account. The cost per oligo is an easy way to compare different companies, however care should be taken to compare like for like. If an amino modified group is to be used for the attachment of the oligo to the slide ensure this is included in the price. The approximate number of probes that are to be designed also needs to be decided on before deciding on a supplier as oligos are supplied in 96-well or 384-well plates with the price often calculated per plate. Therefore, it is often a requirement to fill at least 90% of a plate or an extra cost will be incurred. The flexibility of what can be synthesised is also variable between man-

ufacturers, some suppliers insist that all oligos in a plate must be of the same length; this will obviously effect the design strategy.

15. On a practical level, it is necessary to minimise any bias that may be inadvertently introduced into the experiment, this includes the extraction of RNA. For the entire time course 10 samples are to be taken from each of the three biological replicates. This will result in a total of 30 samples for which RNA needs to be extracted. It is not possible to extract all these samples on a single day; therefore, the process was split over 2 days. Avoid instances such as all the samples extracted on day 1 being from the first five time points. Equally if the process of RNA extraction is split between two people, one person should not extract all samples from a single biological replicate.
16. Some of the most common are: poly-L-lysine, which forms ionic interactions with phosphate of DNA, gamma amino propyl silane (GAPS), which also interacts with phosphate of the DNA, epoxide/silane surfaced slides where an amino modification of the DNA forms an covalent bond with the surface, and aldehyde-based slides, which allow the formation of Schiff bonds.
17. If the probes that were designed were sense, it is possible to incorporate aa-dUTP at this point. This will allow the aRNA to be hybridised to the array. The disadvantage of this is that the aRNA is more labile than DNA.
18. It is recommended by Ambion that a hybridisation oven is used for this step. However, we found that same results were obtained whilst using a thermocycler.
19. Indirect labelling was used as it is thought to give better sensitivity, reduced dye biases and decreased cost compared to direct labelling *(46)*. However, it does take longer than a direct method of incorporation.
20. The buffer provided by Abgene was not used as this contains Tris. If all traces of Tris are not removed then the monofunctional NHS-ester Cy dyes can combine with free amine groups in solution. To avoid this, the 80% ethanol which is tris-free was used as an alternative.
21. The NHS-esther on the Cy dye binds to the amine modification of the cDNA. The Cy dye allows visualisation and quantification of the amount of hybridisation between probe and target. If un-incorporated Cy dye is not removed, there will be a high level of background fluorescence.
22. The choice of labelling the control sample with Cy3 and the experimental sample with Cy5, or vice versa, is technically an arbitrary decision. However, it is most common to label samples this way and some analysis software assumes

this direction. It is easy to analyse data dyed in the 'opposite' direction, but care should be taken to ensure that software assumptions do not affect interpretation of results.

23. The hydroxylamine quenches any uncoupled Cy dye, to prevent both samples being labelled with the same dye.

References

1. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray*. Science, 1995. **270**(5235): pp. 467–70.
2. Luke, K., et al., *Microarray analysis of gene expression during bacteriophage T4 infection*. Virology, 2002. **299**(2): pp. 182–91.
3. Duplessis, M., et al., *Global gene expression analysis of two Streptococcus thermophilus bacteriophages using DNA microarray*. Virology, 2005. **340**(2): pp. 192–208.
4. Frye, J.G., et al., *Host gene expression changes and DNA amplification during temperate phage induction*. J Bacteriol, 2005. **187**(4): pp. 1485–92.
5. Clokie, M.R., et al., *Transcription of a 'photosynthetic' T4-type phage during infection of a marine cyanobacterium*. Environ Microbiol, 2006. **8**(5): pp. 827–35.
6. Kohane, I.S., A.J. Kho, and A. Butte, *Integrative Genomics*. 2003: MIT Press.
7. Stekel, D.M.B.I.-.-.-X., *Microarray Bioinformatics*. 2003: Cambridge University Press. 280.
8. Tiquia, S.M., et al., *Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples*. Biotechniques, 2004. **36**(4): pp. 664–675.
9. Hughes, T.R., et al., *Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer*. Nat Biotechnol, 2001. **19**(4): pp. 342–7.
10. He, Z., et al., *Use of microarrays with different probe sizes for monitoring gene expression*. Appl Environ Microbiol, 2005. **71**(9): pp. 5154–62.
11. Shoemaker, D.D., et al., *Experimental annotation of the human genome using microarray technology*. Nature, 2001. **409**(6822): pp. 922–7.
12. Kane, M.D., et al., *Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays*. Nucleic Acids Res, 2000. **28**(22): pp. 4552–7.
13. Barczak, A., et al., *Spotted long oligonucleotide arrays for human gene expression analysis*. Genome Res, 2003. **13**(7): pp. 1775–85.
14. Wang, H.Y., et al., *Assessing unmodified 70-mer oligonucleotide probe performance on glass-slide microarrays*. Genome Biol, 2003. **4**(1): p. R5.
15. Zhang, W., I. Shmulevich, and J. Astola, *Microarray quality control*. 2004, Hoboken, N.J.: Wiley-Liss. xii, 136.
16. Mann, N.H., et al., *The genome of S-PM2, a "photosynthetic" T4-type bacteriophage that infects marine Synechococcus strains*. J Bacteriol, 2005. **187**(9): pp. 3188–200.
17. Miller, E.S., et al., *Bacteriophage T4 genome*. Microbiol Mol Biol Rev, 2003. **67**(1): pp. 86–156, table of contents.
18. Koehler, R.T. and N. Peyret, *Effects of DNA secondary structure on oligonucleotide probe binding efficiency*. Comput Biol Chem, 2005. **29**(6): pp. 393–7.
19. He, Z., et al., *Empirical establishment of oligonucleotide probe design criteria*. Appl Environ Microbiol, 2005. **71**(7): pp. 3753–60.
20. Rhee, S.K., et al., *Detection of genes involved in biodegradation and biotransformation in microbial communities by using 50-mer oligonucleotide microarrays*. Appl Environ Microbiol, 2004. **70**(7): pp. 4303–17.
21. Taroncher-Oldenburg, G., et al., *Oligonucleotide microarray for the study of functional gene diversity in the nitrogen cycle in the environment*. Appl Environ Microbiol, 2003. **69**(2): pp. 1159–71.
22. Letowski, J., R. Brousseau, and L. Masson, *Designing better probes: effect of probe size, mismatch position and number on hybridization in DNA oligonucleotide microarrays*. J Microbiol Methods, 2004. **57**(2): pp. 269–78.
23. Chung, W.H., et al., *Design of long oligonucleotide probes for functional gene detection in a microbial community*. Bioinformatics, 2005, **21**(22): pp. 4092–100.
24. Nygaard, V. and E. Hovig, *Options available for profiling small samples: a review of sample amplification technology when combined with microarray profiling*. Nucleic Acids Res, 2006. **34**(3): pp. 996–1014.
25. Puskas, L.G., et al., *RNA amplification results in reproducible microarray data with*

slight ratio bias. Biotechniques, 2002. **32**(6): pp. 1330–4, 1336, 1338, 1340.
26. Jenson, S.D., et al., *Validation of cDNA microarray gene expression data obtained from linearly amplified RNA*. Mol Pathol, 2003. **56**(6): pp. 307–12.
27. Zhu, B., F. Xu, and Y. Baba, *An evaluation of linear RNA amplification in cDNA microarray gene expression analysis*. Mol Genet Metab, 2006. **87**(1): pp. 71–9.
28. Wang, E., et al., *High-fidelity mRNA amplification for gene profiling*. Nat Biotechnol, 2000. **18**(4): pp. 457–9.
29. Kaposi-Novak, P., et al., *Oligonucleotide microarray analysis of aminoallyl-labeled cDNA targets from linear RNA amplification*. Biotechniques, 2004. **37**(4): pp. 580, 582–6, 588.
30. Hu, L., et al., *Obtaining reliable information from minute amounts of RNA using cDNA microarrays*. BMC Genomics, 2002. **3**(1): p. 16.
31. Baugh, L.R., et al., *Quantitative analysis of mRNA amplification by in vitro transcription*. Nucleic Acids Res, 2001. **29**(5): p. E29.
32. Wang, E., *RNA amplification for successful gene profiling analysis*. J Transl Med, 2005. **3**: p. 28.
33. Nygaard, V., et al., *Effects of mRNA amplification on gene expression ratios in cDNA experiments estimated by analysis of variance*. BMC Genomics, 2003. **4**(1): p. 11.
34. Hessner, M.J., et al., *Use of a three-color cDNA microarray platform to measure and control support-bound probe for improved data quality and reproducibility*. Nucleic Acids Res, 2003. **31**(11): p. e60.
35. Zhang, L., T. Hurek, and B. Reinhold-Hurek, *Position of the fluorescent label is a crucial factor determining signal intensity in microarray hybridizations*. Nucleic Acids Res, 2005. **33**(19): p. e166.
36. Smyth, G.K., J. Michaud, and H.S. Scott, *Use of within-array replicate spots for assessing differential expression in microarray experiments*. Bioinformatics, 2005. **21**(9): p. 2067–75.
37. Scholtens, D. and A. von Heydebreck, *Analysis of differential gene expression studies*, in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, R. Gentleman, et al., Editors. 2005, Springer: New York. pp. 229–248.
38. Cleveland, W.S., *Robust locally weighted regression and smoothing scatterplots*. J Am Stat Assoc, 1979. **74**: pp. 829–836.
39. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. Nucleic Acids Res, 1997. **25**(17): pp. 3389–402.
40. Chou, H.H., et al., *Picky: oligo microarray design for large genomes*. Bioinformatics, 2004. **20**(17): pp. 2893–902.
41. Nordberg, E.K., *YODA: selecting signature oligonucleotides*. Bioinformatics, 2005. **21**(8): pp. 1365–70.
42. Kerr, M.K. and G.A. Churchill, *Experimental design for gene expression microarrays*. Biostatistics, 2001. **2**(2): pp. 183–201.
43. Yang, Y.H. and T. Speed, *Design issues for cDNA microarray experiments*. Nat Rev Genet, 2002. **3**(8): pp. 579–88.
44. Peixoto, B.R., et al., *Evaluation of reference-based two-color methods for measurement of gene expression ratios using spotted cDNA microarrays*. BMC Genomics, 2006. **7**: p. 35.
45. Millard, A., et al., *Genetic organization of the psbAD region in phages infecting marine Synechococcus strains*. Proc Natl Acad Sci U S A, 2004. **101**(30): pp. 11007–12.
46. Holloway, A.J., et al., *Options available–from start to finish–for obtaining data from DNA microarrays II*. Nat Genet, 2002. **32 Suppl**: pp. 481–9.
47. SantaLucia, J., Jr., *A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics*. Proc Natl Acad Sci U S A, 1998. **95**(4): pp. 1460–5.
48. Dudley, A.M., et al., *Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range*. Proc Natl Acad Sci U S A, 2002. **99**(11): pp. 7554–9.
49. Kim, H., et al., *Use of RNA and genomic DNA references for inferred comparisons in DNA microarray analyses*. Biotechniques, 2002. **33**(4): pp. 924–30.
50. Page, G.P., et al., *The PowerAtlas: a power and sample size atlas for microarray experimental design and research*. BMC Bioinformatics, 2006. **7**: p. 84.
51. Guo, Z., et al., *Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports*. Nucleic Acids Res, 1994. **22**(24): pp. 5456–65.
52. LaForge, K.S., et al., *Detection of single nucleotide polymorphisms of the human mu opioid receptor gene by hybridization or single nucleotide extension on custom oligonucleotide gelpad microchips: potential in studies of addiction*. Am J Med Genet, 2000. **96**(5): pp. 604–15.
53. Burgoon, L.D., et al., *Protocols for the assurance of microarray data quality and process control*. Nucleic Acids Res, 2005. **33**(19): p. e172.

54. Taylor, E., et al., *Sequence verification as quality-control step for production of cDNA microarrays*. Biotechniques, 2001. **31**(1): pp. 62–5.
55. t Hoen, P.A., et al., *Intensity-based analysis of two-colour microarrays enables efficient and flexible hybridization designs*. Nucleic Acids Res, 2004. **32**(4): p. e41.
56. Van Gelder, R.N., et al., *Amplified RNA synthesized from limited quantities of heterogeneous cDNA*. Proc Natl Acad Sci U S A, 1990. **87**(5): pp. 1663–7.

Chapter 13

Purification of Bacteriophages and SDS-PAGE Analysis of Phage Structural Proteins from Ghost Particles

Pascale Boulanger

Abstract

Concentration and purification of infectious particles are prerequisites for structural and functional characterization of bacteriophages. The methods detailed in the first part of this chapter outline the protocols commonly used to obtain purified phages: the concentration of phage particles by precipitation with polyethylene glycol and their purification by centrifugation in CsCl step gradients and subsequently by equilibrium centrifugation. This sequence of procedures, if carried out as a whole, ensures a purification of high quality, which is well suited for most analytical techniques used to characterize bacteriophage particles.

The second part of this chapter describes the preparation of "ghosts" or DNA-less bacteriophages. These particles should be preferred to the entire bacteriophages for one-dimensional SDS-PAGE analysis of phage structural proteins, since running of the phage proteins through the gel is not disturbed by the presence of the phage DNA. This allows an optimal resolution, which is necessary for proteomic approaches such as *N*-terminal protein sequencing or mass spectrometry using proteins isolated from distinct gel bands.

Key words: Bacteriophage, purification, polyethylene glycol, CsCl, step gradient, equilibrium centrifugation, SDS-PAGE, ghosts.

1 Introduction

The renewed interest for bacteriophages during the last decade has promoted development of new methods for bacteriophage study having stringent requirements with respect to phage sample quality and purity. Nowadays, the main fields for which purified bacteriophages are necessary, include morphological characterization of viruses by three-dimensional image reconstruction of phage structures with the help of high resolution cryo-electron

Martha R. J. Clokie, Andrew M. Kropinski (eds.), *Bacteriophages: Methods and Protocols, Volume 2: Molecular and Applied Aspects, vol. 502*, © 2009 Humana Press, a part of Springer Science+Business Media
DOI 10.1007/978-1-60327-565-1_13 Springerprotocols.com

microscopy, genomic analysis involving phage DNA extraction for genome sequencing and annotation (**Chapter 22**), as well as proteomic approaches to the identification of viral structural proteins (**Chapter 35**). The combination of these recent developments permits major advancements in the understanding of phage structures. The subnanometer resolution of phage particle architectures, associated with the knowledge of crystallographic structures of phage proteins, allows morphological comparisons encompassing a wide variety of bacteriophages. This provides extensive structural information that sometimes unmasks evolutionary relationships which may be obscured by genome or protein sequence comparisons *(1, 2, 3, 4)*.

Lytic phage progeny are harvested after they have been released from their host bacteria through the process of cell lysis. The major components that are present in crude phage cultures, apart from phage themselves, are bacterial debris—mainly membranes—with bacterial proteins, nucleic acids and ribosomes. In the case of lysates from Gram-negative bacteria, debris also include endotoxins or lipopolysaccharides (LPS), a major component of the bacterial outer membrane which is released in the cultures upon lysis. For most biochemical and biophysical approaches which aim to characterize the bacteriophage structures, a careful removal of all these contaminants is essential in order to obtain highly purified phage preparations. Furthermore, the recent return to bacteriophage-based antibacterial therapy purpose, or "phage therapy," requires a perfect characterization of the phage preparations. These must be guaranteed to have a very low level of contaminants, especially of endotoxins that are highly toxic for humans and animals.

2 Materials

2.1 Phage Purification

1. Phage suspension buffer also called TM buffer (Tris-Mg^{2+} Buffer) 10 mM Tris–HCl (pH 7.2–7.5), 100 mM NaCl, 10 mM $MgCl_2$. Addition of 1–10 mM $CaCl_2$ in the suspension buffer may be required for the stability of some phages.
2. DNase I and RNase A from Bovine Pancreas (Roche or Calbiochem). Stock solutions 1 mg/mL are stored at −20 °C.
3. Chloroform.
4. Sodium chloride powder: NaCl ≥ 99.5 % for molecular biology.
5. Polyethylene glycol powder: PEG 6,000 (MW 5,000–7,000 g/mol) for molecular biology and biochemical purposes.
6. Cesium chloride: CsCl ≥ 99.9 % for density gradient purification.

7. Ultracentrifuge equipments: Beckman L8-55M or equivalent.
8. Swinging-bucket rotors (Beckman): SW41 or SW28 and SW50 or SW65.
9. Centrifuge tubes (Beckman): thinwall or thickwall open-top polyallomer tubes:
   13.2 mL, 14 mm × 89 mm for the SW41 rotor.
   38.5 mL, 25 mm × 89 mm for the SW28 rotor.
   5.0 mL, 13 mm × 51 mm for the SW50 or SW65 rotors.
10. Syringes and 18–22 gauge hypodermic needles.
11. Dialysis tubing: Spectra/Por molecular-porous membrane tubing, MWCO 12–14,000.
12. Refractometer (optional).

2.2 Ghost Preparation

1. Lithium chloride: LiCl ≥ 99. 0 % GR for analysis (Merck) or equivalent.
2. CryoTubes Vials: 3.6 mL or 1.8 mL (Nunc).
3. Liquid nitrogen or ice-ethanol freezing bath.
4. Deoxyribonuclease I: RNase-free DNase I, 5–10U/μL (GE Healthcare) or equivalent.
5. Protease inhibitor cocktail tablets: Complete EDTA-free (Roche Diagnostics GmbH).

2.3 SDS-PAGE Analysis

1. SDS-PAGE equipment: mini format electrophoresis apparatus for 6–10 cm length gels.
2. Handcast gels of variable acrylamide percentage or ready to use precast linear gradient gels (Tris–HCl from Bio-Rad or NuPAGE Novex from Invitrogen).
3. Standard Laemmli loading buffer and running buffer.
4. Standard solutions for Silver or Coomassie blue staining. The ready to use solution Bio-Safe Coomassie G-250 (Bio-Rad) or SimplyBlue SafeStain (Invitrogen) that does not require ethanol and acetic acid for destaining may also be used.

3 Methods

3.1 Concentration of Bacteriophages by Precipitation with Polyethylene Glycol (PEG)

Bacteriophages can be concentrated from crude lysates of infected bacteria after addition of polyethylene glycol. This method, whose efficiency is almost independent of phage concentration, is useful in order to concentrate phages even with very low titer lysates *(5)*. It is a mild and fast procedure allowing a 100-fold phage concentration after low speed centrifugation with negligible loss of infectivity. Furthermore, it is applicable to most bacteriophages without modification.

1. Add DNase I and RNase A to 1 μg/mL each to the bacterial lysate freshly recovered from phage growth in liquid cultures or from soft-agar overlay cultures (*see* **Chapter 7**).

This step completes degradation of residual bacterial DNA and unpackaged phage DNA and permits dissociation of ribosomes which could contaminate phage preparations (*see* **Note 1**). At the same time, add 0.2% (v/v) chloroform to complete lysis (chloroform addition should be omitted for lipid-containing phages) and incubate the preparation for 30 min at room temperature.

2. Dissolve solid NaCl into the bacteriophage suspension to the concentration of 0.5 M and let it cool at 4 °C for 1 h. NaCl promotes dissociation of phage particles from bacterial debris and is required for the next step of precipitation with polyethylene glycol.
3. Remove the denser bacterial debris by centrifugation of the suspension at 6,000–8,000 g for 10 min at 4 °C. Then transfer the phage-containing supernatant into a clean flask.
4. Dissolve PEG 6000 to a final concentration of 8%−10% at 4 °C, by brief stirring, and let it stand at 4 °C for at least 1 h in order to precipitate phage particles (*see* **Note 2**).
5. Sediment the precipitated phage at 10,000 *g* for 15 min at 4 °C and carefully discard the supernatant. The pellet forms a film, which sticks to the wall of the centrifuge bottles. Turn the centrifuge bottles over and let the remaining fluid drain away from the pellet for 5 min.
6. Gently suspend the pellet in phage suspension buffer (1–2 ml per 100 ml of supernatant). Since phage particles may be damaged by vortexing or vigorous pipetting, it is recommended to use a wide-bore pipette equipped with a bulb or an automatic pipette equipped with a truncated tip (1 or 5 ml). Alternatively, the pellet can be left overnight at 4 °C in order to soften, which facilitates the suspension.
7. Separate phage particles from co-precipitated bacterial debris by low-speed centrifugation for 10 min at 5000 g, at 4 °C.
8. If it is not needed to go any further in the purification, the residual PEG and bacterial debris can be removed. This may be done by gentle extraction for 1 min with an equal volume of chloroform. The phage containing aqueous phase is separated from the white organic phase by centrifugation at 5,000 g for 15 min.

The above procedure of concentration with PEG is in itself a purification method which can be sufficient depending on the grade of purity required for the phage preparation. Actually, this procedure yields a partially purified preparation useful for several applications. If desired, phages can be further purified by centrifugation in a CsCl gradient.

3.2 CsCl Purification of Concentrated Bacteriophages

Phage particles can be separated from contaminants according to their buoyant density by sedimentation in CsCl gradients. Highly pure and concentrated phage preparations are obtained after two

successive centrifugations in CsCl gradients: i) A centrifugation on preformed CsCl step gradients for a rapid discarding of most debris and contaminants. ii) An equilibrium centrifugation (isopycnic sedimentation) through a self-generating CsCl gradient ensuring a careful and complete elimination of the residual contaminants. This method is applicable to most bacteriophages, provided that their stability in the presence of high concentrations of CsCl has been checked.

Although it remains the most widely used procedure to achieve a high degree of purification, some bacteriophages may also be purified by centrifugation on glycerol or sucrose gradients or by ion exchange chromatography (*see* **Note 3**).

3.2.1 Gradient Step Centrifugation

The centrifugation in step gradients is suitable for the purification of large-scale preparations. It should be performed in large centrifuge tubes adapted to the Beckman swinging-bucket rotors SW41 or SW28 (or equivalent centrifuge device), depending on the volume of phage suspension to be purified (*see* **Note 4**). It provides quite pure phage particles suspensions that are usable for many applications.

Bacteriophages, usually recovered from precipitation with PEG, are sedimented through a discontinuous step gradient formed from successive layers of CsCl solutions of increasing density. The densities of the different CsCl layers are chosen so that the density range encompasses the proper buoyant density of the phage. If this latter value is unknown or if it cannot be estimated from the phage physical characteristics (*see* **Note 5**), it may be necessary to test several CsCl layer density patterns to optimize the purification.

1. Prepare the different CsCl solutions by dissolving the salt in the phage buffer. The most commonly used solutions are listed in **Table 13.1** and a reliable method to prepare a solution of a given density is detailed in **Note 6**. The density of each solution can be checked simply by weighting 1 ml of it on a precision weighting scale or, more accurately, by measuring the refractive index if a refractometer is available.
2. Prepare step gradients by hand, using a manual precision pipette. Carefully layer solutions of decreasing density by placing the tip of the pipette at the angle formed by the tube wall and the meniscus. Alternatively, you may float the lighter gradient concentrations up by adding increasing density solutions to the tube bottom using a hypodermic syringe with a long needle.
3. Once the gradient is ready, carefully layer the phage suspension on the top of the gradient. In case of using thinwall centrifuge tubes, mind to fill the tubes to within 3–5 mm of the top for proper tube support. So it can be necessary

Table 13.1
CsCl solutions currently used for phage purification

d (g/mL)	*c* (g/mL)	*c* (g/50 mL)	*n* (refractive index)
1.20	0.275	13.74	1.3527
1.25	0.342	17.11	1.3575
1.30	0.410	20.49	1.3622
1.40	0.546	27.28	1.3717
1.45	0.614	30.70	1.3765
1.50	0.683	34.13	1.3813
1.60	0.820	41.02	1.3908
1.70	0.959	47.96	1.4003

to add phage buffer to the phage sample if the tube is partially filled. If a thickwall tube is used, additional buffer is not required.

4. Centrifuge at a rotor speed of 22,000–25,000 rpm (relative centrifugal field of 100,000–120,000 g) for 2–3 h, in a Beckman SW41 or SW28 rotor (or equivalent).
5. After centrifugation you will observe the following distribution: the majority of the bacterial debris, especially membranes and endotoxins, whose density ranges from 1.15 to 1.25 *(6)* do not sediment further than the 1.3 density layer. The phage particles sediment through the lowest density layers until they reach the layer whose density is equal or greater than their proper buoyant density. Thus tailed phages form a bluish-white and opalescent band located at the interface between the 1.4 and 1.5 or between the 1.5 and 1.6 density layers. The example of the purification of bacteriophage T5 is shown in **Fig. 13.1-A**.
6. The phage band can be simply collected from above. Carefully remove all the upper layers containing the contaminants and debris by using a micropipette or a Pasteur Pipette. Then, with a clean tip placed just underneath the target band, collect the phage particles. When thinwall tubes are used, you can alternatively puncture the tube wall with a needle connected to a syringe, just below the desired band, and slowly aspirate the phages.
7. Remove the CsCl from the phage suspensions by dialysis at 4 °C, two or three times for 30 min against ca. 500 volumes of phage buffer or overnight against 2,000 volumes of phage buffer. Store the dialyzed phage suspension at 4 °C with a few drops of chloroform to avoid microbial contamination if the

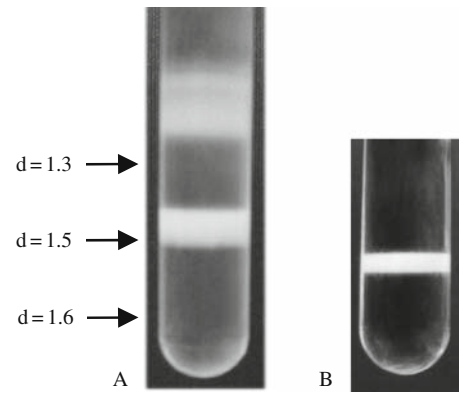


Fig. 13.1. Purification of bacteriophage T5 by CsCl centrifugation. (**A**) CsCl step gradient: bacteriophage T5 (3 mL) harvested after precipitation with PEG, were layered over a CsCl step gradient preformed in a Thinwall Pollyallomer tube (No. 331372 from Beckman-Coulter). The CsCl layers were $d = 1.6$:2 mL, $d = 1.5$:3 mL, $d = 1.3$:3 mL. After centrifugation in a rotor SW41 at $4\,^{\circ}C$, for 2 hr 30 min at 25,000 rpm, the debris formed a white-yellow zone above the layer of density 1.3 and the phage particles formed an opalescent band above the layer of density 1.5 (**B**) Equilibrium centrifugation: 2.5 mL of phage T5 were collected from the step gradient (**A**). The measured refractive index was $n = 1.3805 \pm 0.0002$ indicating the density 1.495. The volume was adjusted to 5 mL with the 1.5 density solution in an Ultra-Clear Thinwall Pollyallomer tube (No. 344057 from Beckman-Coulter) and centrifuged for 22 h, at 38,000 rpm and $4\,^{\circ}C$ in a SW65 rotor.

phage tolerates chloroform. Other alternative storage conditions are detailed in **Chapter 15**.

8. For proteomic applications such as N-terminal protein sequencing or mass spectrometry, which aim to identify the phage structural proteins, contaminating host proteins or endotoxins should be strictly removed to avoid background signals (*see* **Chapter 34**). For this purpose, an equilibrium centrifugation in CsCl is recommended. If this is not possible, a second gradient step centrifugation can be carried out with a dialysis or a slow dilution step of the phage suspension between the two centrifugation runs.

3.2.2 Equilibrium Centrifugation in CsCl

This method is commonly used as the final purification step of phage preparations, but it can also be used as a single purification method after the precipitation with PEG when dealing with small-scale preparations of bacteriophage.

Centrifugation of a homogeneous density CsCl solution generates a continuous density gradient. The density of the starting solution and the rotor speed are selected so that at equilibrium, the density range from the top to the bottom of the gradient is sufficient to encompass buoyant densities of the particles to be separated, i.e., debris and bacteriophages. Banding of tailed bacteriophages can be obtained with a starting solution at the

density 1.5 g/mL by using 5 mL capacity centrifuge tubes that fit the swinging-bucket rotors SW50 or SW65 (Beckman).

1. If the phage suspension does not contain any CsCl, adjust its density to 1.5 g/mL by dissolving 0.75 g of solid CsCl per ml of suspension. CsCl should be added gradually to prevent inactivation of bacteriophages by osmotic shock.
2. If the bacteriophages have been previously purified by a step gradient centrifugation, the density of the phage suspension must be already close to 1.5. Check the exact density by measuring the refractive index (1.3813 ± 0.0005) or by weighting the suspension, and adjust to the density 1.5 if necessary.
3. Adjust the final volume to 5 mL with a 1.5 g/mL CsCl solution and transfer the phage suspension to ultracentrifuge tubes that fit the swinging-bucket rotor.
4. Centrifuge at 35,000 rpm (SW50) or 38,000 rpm (SW65) for 18–24 h at 4 °C. This corresponds to a RCF of 150,000 g.
5. The purified phage particles form a band equilibrated at a position corresponding to their proper buoyant density (**Fig. 13.1.B**). Collect the band and dialyze the harvested suspension against phage buffer as described in **Section 3.2.1**.

3.3 Preparation of Bacteriophage Ghosts for 1D Gel Electrophoresis Analysis of Structural Phage Proteins

One-dimensional sodium dodecylsulfate-polyacrylmamide gel electrophoresis (1D SDS-PAGE) is the most suitable method for a first analysis of the phage structural proteins from purified phage particles. It helps to identify the structural protein genes once the phage's genome has been sequenced. Indeed, *N*-terminal protein sequencing using Edman degradation as well as mass spectrometry of the phage structural proteins are usually carried out from protein bands separated by 1D gel electrophoresis.

Analysis of phage proteins by SDS-PAGE can be made from entire phage particles or from DNA-less particles or "ghosts." Phage particles can be used directly if all structural proteins can be located on the gels, even from not very concentrated samples. However, some minor structural proteins may be difficult to visualize, unless large amount of phage samples have been loaded. In this latter case, the phage DNA contained in the concentrated phage samples would interfere with the protein separation and weaken the resolving power of the electrophoresis. For this reason, analysis of ghost preparations is recommended, especially for large-genome bacteriophages (dsDNA greater than 50 kb). Described below are two methods of ghost production: the treatment of phage with lithium chloride and the freeze–thawing treatment. Alternative methods have been described that are not detailed in this section because of their lower yield (*see* **Note** 7).

3.3.1 Preparation of DNA-Less Ghosts

3.3.1.1 Treatment of Bacteriophages with Lithium Chloride

1. Mix 1 volume of purified phage particles (from 1×10^{11} to 5×10^{12} pfu/mL) with an equal volume of a 10 M LiCl solution, and incubate for 10 min at 46 °C. The suspension becomes more or less viscous, depending on the phage concentration and on the size of the phage DNA.
2. Dilute the mixture 10-fold with phage suspension buffer.
3. Add 10 mM $MgCl_2$ (or $MgSO_4$) and 50 U of RNase-free DNaseI per 1×10^{12} pfu, and incubate at 37 °C for 2 h (*see* **Note 8**).
4. Concentrate the ghost particles by ultracentrifugation at 100,000 g, 4 °C for 30 min. This may be done at 32,000 rpm with a 45Ti rotor in a classical ultracentrifuge for volumes until 70 mL, or at 50,000 rpm with a TL100.3 rotor in a benchtop ultracentrifuge for volumes until 3 mL (from Beckman or equivalent).
5. Let the ghost pellet suspend in phage buffer. At this stage, the ghost suspension may be concentrated with regard to the initial phage suspension, in order to facilitate the detection of minor proteins by SDS-PAGE.

3.3.1.2 Freeze–Thawing of Bacteriophages

1. The titer of the bacteriophage suspension should not exceed 5×10^{11} pfu/mL, especially for phages with DNA larger than 100 kb. It is recommended to use cryotubes and to treat volumes that do not exceed 3 mL.
2. Freeze the phage suspension in liquid Nitrogen (or in an ice-ethanol freezing bath providing a temperature below −10 °C) and immediately thaw it in a water bath at 46 °C. Repeat this at least four times.
3. Add 10 mM $MgCl_2$ (or $MgSO_4$) if not included in the phage suspension.
4. Incubate the ghost suspension with DNaseI and concentrate the particles as detailed in **Section 3.3.1.1**

3.3.2 1D Gel Electrophoresis of Phage and Ghost Particles

1. Suspend the phage or ghosts particles in Laemmli loading buffer containing 100 mM β-mercaptoethanol or dithiothreitol. The amount of particles should be adapted to each type of bacteriophage.
2. Denaturate the phage or ghost samples by heating at 100 °C for 5 min. DNA from entire phage particles or residual DNA from some ghost preparation may render the samples viscous after heating, thus bothering loading onto the gels. Repeated pipetting of the hot sample by using a thin tip followed by prolonged denaturation may help loading. If this treatment is not sufficient, it may be necessary to dilute the sample.
3. Resolve the proteins by using adapted SDS-PAGE gels. If possible, use gradient gels which allows resolving proteins within a large range of molecular weight (**Fig. 13.2**).

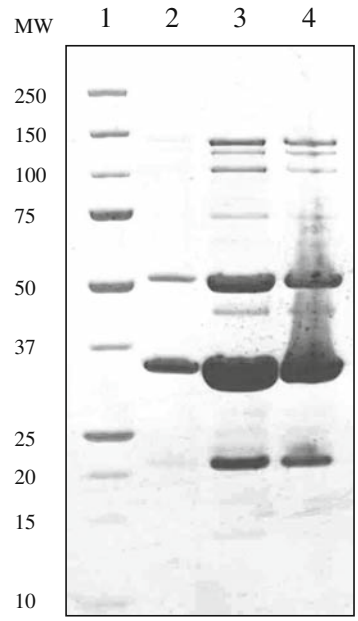


Fig. 13.2. SDS-PAGE analysis of T5 bacteriophage and ghost particles. The structural proteins were resolved in a NuPAGE 4–12% bis-tris gel in MES–SDS running buffer (Invitrogen) and stained with Bio-Safe Coomassie G-250 (Bio-Rad) 1: Precision Plus Protein Standards Bio-Rad. 2: Bacteriophage T5: 1×10^{10} pfu. 3: Ghosts prepared by LiCl treatment: 2.5×10^{11} particles. 4: Ghosts prepared by freeze–thawing: 2.5×10^{11} particles. All samples were denaturated in $20\,\mu L$ of Laemmli loading buffer for 5 min at $100\,^{\circ}C$, or for extended time if necessary. The amount of phage T5 that could be loaded on the gel was twenty times less than the amount of T5 ghosts. Residual DNA which is included in the ghost preparation obtained by freeze–thawing decreases the resolution of the T5 structural proteins. This is often observed with large genome bacteriophages.

4 Notes

1. Most of the time bacterial DNA is degraded by phage encoded endonucleases and contamination of crude lysates by bacterial DNA is limited. Ribosomes sediment at PEG concentrations over 5% and are pelleted at low speeds *(5)*. The contamination with ribosomes is drastically decreased by a treatment of phages lysates with RNAse at the concentration of $1\,\mu g/mL$.
2. Although different phage may require different concentration of PEG for maximum efficiency of precipitation, the PEG concentration of 10% allows at least 90% of the infective titer of most phage to be pelleted in 1 h at $4\,^{\circ}C$. For small bacteriophage whose density is less than 1.4, longer periods of standing at $4\,^{\circ}C$ prior to centrifugation may increase the fraction of phage particles found in the pellet *(5)*.

3. Some bacteriophages may be purified by a rate zonal centrifugation in sucrose gradient, followed by an equilibrium centrifugation in a sucrose gradient. This procedure is well suited to the purification of lipid containing phages, whose density is close to 1.3 g/mL *(7)*.

   It should be noted that some bacteriophage may be damaged by the centrifugations in CsCl or sucrose gradients *(7)*. An alternative method, which utilizes anion exchange chromatography on commercially available cartridges, has been described for the purification of the lipid-containing PRD1 bacteriophage. This procedure proved to have the advantage of preserving the integrity and infectivity of the phage particles *(8)*.
4. Centrifuge tubes: see http://www.beckmancoulter.com for exhaustive specification of rotors and tubes.
5. The data available for phage that have been identified so far give an overview of the buoyant density of bacteriophages from different families (*see* **ref**. *(9)* for a catalog of physical properties of different phages and **ref**. *(10)* for a recent review on the phage classification or **Volume 1 Chapter 13** in this book). Tailed phages, which represent the vast majority of phages (96%), have a buoyant density between 1.45 and 1.52. Other phages (4%) including polyhedral, filamentous, and pleomorphic phages have a buoyant density comprised between 1.27 and 1.47, depending on their characteristics: close to 1.3 for lipid containing phages, close to 1.4 for others.
6. For the preparation of CsCl solutions at a given density d (g/mL), we recommend to use the following formula to calculate the final CsCl concentration c (g/mL): $c = 0.0478\ d^2 + 1.23\ d - 1.27$. The corresponding refractive index n can be calculated from $n = 0,\ 0951\ d + 1,\ 2386$

   The above formula were established from the concentrative properties of aqueous CsCl solutions, as referenced in the Handbook of Chemistry and Physics *(11)*, and available online at http://www.hbcpnetbase.com. They are valid for solutions whose density is comprised between 1.0 and 1.9 g/mL.
7. Osmotic shocks of phages *(12, 13, 14)*, as well as treatment with the chelating agent EDTA *(12, 15)* have also been used for ghost preparation. However, the yields obtained by these methods, which were used for the preparation of bacteriophage T5 ghosts (dsDNA 121,750 bp) were less than 20% (Bonhivers, M., (1995) PhD Thesis, unpublished results).
8. Digestion of the phage DNA released upon ghost formation should be done with highly pure DNase, guaranteed free of protease contaminants which could degrade some phage structural proteins. The commercially available RNase-free DNAse solutions are recommended. Otherwise, protease inhibitor cocktail tablets may be added to the ghost suspension during the treatment with DNase.

References

1. Bamford, D.H., Grimes J.M. and Stuart, D.I. (2005) What does structure tell us about virus evolution? *Curr. Opin. Struct. Biol.* **15**, 655–663.
2. Effantin, G., Boulanger, P., Neumann, E., Letellier, L. and Conway, J. F. (2006) Bacteriophage T5 structure reveals similarities with HK97 and T4 suggesting evolutionary relatationships. *J. Mol. Biol.* **361**, 993–1002.
3. Jiang,W., Chang J., Jakana,J., Weigele, P., King, J. and Chiu, W. (2006) Structure of epsilon15 bacteriophage reveals genome organization and DNA packaging/injection apparatus *Nature*, **439**, 612–616.
4. Fokine, A., Kostyuchenko, V. A., Efimov, A. V., Kurochkina, L. P., Sykilinda, N. N., Robben, J., Volckaert, G., Hoenger, A., Chipman, P. R., Battisti, A. J., Rossmann, M. G., and Mesyanzhinov, V. V. (2005) A three-dimensional Cryo-electron microscopy structure of the bacteriophage ΦKZ head. *J. Mol. Biol.* **352**, 117–124.
5. Yamamoto, K. R. and Alberts, B. M. (1970) Rapid bacteriophage sedimentation in the presence of Polyethylene Glycol and its application to large scale virus purification. *Virolgy* **40**, 734–744.
6. Ishidate, K., Creeger, E. S., Zrike, J., Deb,S., Glauner, B., MacAlister, T.J. and Rothfield, L. I. (1986) Isolation of differentiated membrane domains from Escherichia coli and Salmonella typhimurium, including a fraction containing attachment sites between the inner and outer membranes and the murein skeleton of the cell envelope. *J. Biol. Chem.* **261**, 428–43.
7. Kivela, H.M., Mannisto, R. H., Kalkkinen, N. and Bamford, D.H. (1999) Purification and protein composition of PM2, the first lipid-containing bacterial virus to be isolated. *Virology*, **262**, 364–374.
8. Walin, L., Tuma, R., Thomas, G.R. Jr. and Bamford, D.H. (1994) Purification of viruses and macromolecular assemblies for structural investigations using a novel ion exchange method. *Virology* **201**, 1–7.
9. Fraenkel-Conrat, H. (1985) Phages of prokaryotes (Bacteria and cyanobacteria) in *The viruses. Catalogue, characterization and classification* Plenum Press, New York, pp. 173–222.
10. Ackermann, H. W., (2003) Bacteriophage observations and evolution. *Res. Microbiol.*154, 245–251.
11. Handbook of Chemistry and Physics, 87th edition 2006–2007, CRC press.
12. Konopa, G. and Taylor, K. (1975) Isolation of coliphage lambda ghosts able to adsorb onto bacterial cells. *Biochimica et Biophysica Acta*, **399**, 460–467.
13. Konopa, G and Taylor, K. (1979) Coliphage λ ghosts obtained by Osmotic Shock or LiCl treatment are devoid of J- and H- gene products. *J. Gen. Virol.* **43**, 729–733.
14. Duckworth D. H. (1970) Biological activity of bacteriophages Ghosts and "take over" of host functions by bacteriophage. *Bacteriol. Rev.* **34**, 344–363.
15. Yamamoto, N., Fraser, D. and Mahler, H. R. (1968) Chelating agent shock of bacteriophage T5. *J. Virol.* **2**, 944–950.

Chapter 14

Phage Proteomics: Applications of Mass Spectrometry

Rob Lavigne, Pieter-Jan Ceyssens and J. Robben

Abstract

Current techniques in mass spectrometry (MS) allow sensitive and accurate identification of proteins thanks to the in silico availability of these protein sequences within databases.

This chapter provides a short overview of MS techniques used in the identification of phage structural proteins and focuses on an electron spray peptide ionization (ESI-MS/MS) approach to identify the phage structural proteome in a comprehensive and systematic ways. Such analyses provide an experimental examination of structural proteins and confirm genome-based gene predictions.

Key words: Structural proteome, mass spectrometry, whole-phage shotgun proteomics, bacteriophage.

1 Introduction

In recent years, the use of mass spectrometry (MS) for the identification of structural phage proteins has become increasingly popular and can be considered as a logical next step after phage genome sequencing. Indeed, structural proteins often show poor sequence similarity to proteins in the databases limiting homology-based annotation, and N-terminal sequencing by Edman degradation is expensive and is usually only applicable to major virion proteins. The systematic identification of the structural proteins by MS can provide a more detailed experimental annotation and a confirmation of in silico ORF predictions. Estimates based on comprehensive experimental MS data and genome annotations place the number of encoded structural proteins between 20 and 30% of the total number of phage genes (**Table 14.1**). These observations underline the importance of MS techniques in the identification of the structural proteome.

Martha R. J. Clokie, Andrew M. Kropinski (eds.), *Bacteriophages: Methods and Protocols, Volume 2: Molecular and Applied Aspects, vol. 502*, © 2009 Humana Press, a part of Springer Science+Business Media
DOI 10.1007/978-1-60327-565-1_14 Springerprotocols.com

Table 14.1
List of experimental identifications of structural proteomes from phages

Phage	Genome size	Predicted ORFs	Separation	Identification technique	No. prot ID*	Reference
ϕCTX	35,538	47	1D gel	*N*-terminal sequencing (Edmann degradation)	15	Nakayama et al. (1999) *(1)*
A118	40,834	72	1D gel	*N*-terminal sequencing (Edmann degradation)	2	Loessner et al. (2000) *(2)*
PSA	37,618	57	1D gel	*N*–terminal sequencing (Edmann degradation)/ MALDI-TOF PMF/ESI MS/MS	5	Zimmer et al. (2003) *(3)*
T1	48,836	77	2D gel	MALDI-TOF PMF (Micromass M@LDI R)	4	Roberts et al. (2004) *(4)*
LP65	131,573	165	2D gel	ESI-LC-MS/MS	5	Chibani-Chennoufi et al. (2004) *(5)*
SP6	43,769	52	1D gel	*N*-terminal sequencing (Edmann degradation)	10	Scholl et al. (2004) *(6)*
K1–5	44,385	52	1D gel	*N*-terminal sequencing (Edmann degradation)	10	Scholl et al. (2004) *(6)*
2972	34,704	44	1D gel	*N*-terminal sequencing (Edmann degradation)/ MALDI-TOF PMF (Voyager-DE PRO Biospec. Workstation)	8	Lévesque et al. (2005) *(7)*
F116	65,195	70	1D gel	MALDI-TOF PMF (Micromass M@LDI R)	3	Byrne and Kropinski (2005) *(8)*
BFK20	42,968	55	2D gel	*N*-terminal sequencing (Edmann degradation)	6	Bukovska et al. (2006) *(9)*
ϕKMV	42,519	52	1D gel/	ESI-LC-MS/MS	11	Lavigne et al. (2006) *(10)*
				Peptide ion gas phase fractionation		
LKD16	43,200	54	1D gel/	ESI-LC-MS/MS	13	Ceyssens et al. (2006) *(11)*
				Peptide ion gas phase fractionation		
LKA1	41,593	56	1D gel/	ESI-LC-MS/MS	10	Ceyssens et al. (2006) *(11)*
				Peptide ion gas phase fractionation		

(continued)

Table 14.1 (continued)

Phage	Genome size	Predicted ORFs	Separation	Identification technique	No. prot ID*	Reference
ϕKZ	280,334	306	1D gel/	ESI-LC-MS/MS	62	Lecoutere et al., submitted
				Peptide ion gas phase fractionation		
EL	211,215	201	1D gel/	ESI-LC-MS/MS	64	Lecoutere et al., submitted
				Peptide ion gas phase fractionation		
YuA	58,662	78	1D gel/	ESI-LC-MS/MS	16	Ceyssens (PMID: 18065532)
				Peptide ion gas phase fractionation		
ϕSN	66,391	89	1D gel/	ESI-LC-MS/MS	20	Unpublished data
				Peptide ion gas phase fractionation		

*Number of identified structural proteins

For MS-based identification, proteins are first digested with a specific protease, usually trypsin. The resulting peptide mixture is then analyzed, after ionization, by either electrospray (ESI) or matrix-assisted laser desorption (MALDI). In peptide mass fingerprinting (PMF), the protein is identified by comparison of the peptide ion mass spectrum obtained with the theoretical mass spectra generated from a computer database of protein sequences.

In tandem mass spectrometry (MS/MS), peptide ions are individually mass-selected and further fragmented physically, and the resulting peptide fragmentation spectrum compared with computer-generated fragmentation spectra. In this way, MS/MS spectra allow amino acid sequence-based identification that is more confident than PMF. It is clear that in both cases, the availability of the protein sequences within the database is important for identification. Liquid chromatography (LC) of (tryptic) peptides combined with MS/MS generates large amounts of tandem spectra and allows identification of individual proteins in complex protein samples derived from, e.g., mixed gel bands or even from whole phage particles, circumventing the need of elaborate protein separation prior to identification. Due to the high sensitivity and accuracy of present-day mass spectrometers, also low abundant proteins and potential translational shifts can be identified *(10, 3)*. In addition, MS-based identification of phages as secondary biomarkers for target bacteria has been implemented *(12)*. On the other hand, N-terminal protein sequencing using Edman degradation still may provide additional information on proteolytic cleavage/maturation *(1)*.

Future applications of mass spectrometry may include more complex analyses like the elucidation of the phage infection mechanism and the molecular effects of infection on the host proteome.

In this chapter, we focus on the identification of structural phage proteins. As shown in **Table 14.1**, the structural proteome of a number of phages has been studied by both a 1D gel-based approach and a "whole-phage shotgun analysis" (WSA) approach. In this combined strategy, complementary data are generated which can lead to a more complete coverage of the structural proteome. The flowchart (**Fig. 14.1**) indicates the difference between both approaches.

The different technical steps are boxed, whereas the corresponding results of each step are indicated as dotted boxes. Feedback loop arrows indicate the steps necessary to analyze more samples. For the SDS-PAGE, this means analysis of other gel slices. In WSA, complexity reduction is attained by analyzing different aliquots of the sample in a set of narrow, non-overlapping mass windows.

While the SDS-PAGE approach separates proteins in order to analyze individual protein bands, the (WSA) approach creates a

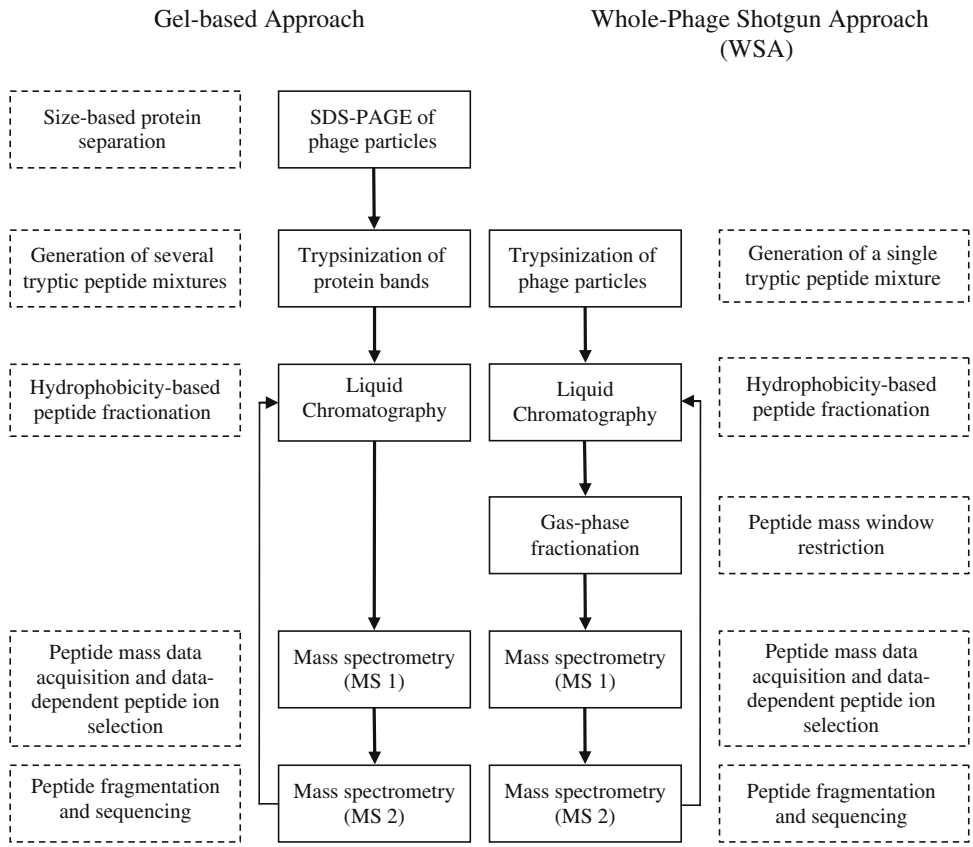


Fig. 14.1. Flowchart of the SDS-PAGE (left) and WSA approach (*right*).

single complex mixture of peptides from all structural phage proteins, which are separated/fractionated in two dimensions (based on polarity and mass) prior to identification. Generally, WSA and peptide fractionation followed by MS/MS can reveal smaller sized and less abundant structural proteins and could be useful for phages that are difficult to prepare.

2 Materials

2.1 Phage Purification and Concentration

1. Buffers and reagents:
   - SM buffer: 1 M Tris–HCl (pH 7.5), 2% (w/v) gelatin, 10 mM $MgSO_4.7H_2O$, 100 mM NaCl
   - Dialysis buffer: 10 mM Tris–HCl (pH 7.5), 100 mM $MgSO_4$, 150 mM NaCl.
   - Four CsCl-solutions with increasing density (**Table 14.2**).

Table 14.2
CsCl-solutions used in phage purification

Density (g/ml)	CsCl (g)	SM buffer (ml)
1.33	11	23.42
1.45	15	21.25
1.50	16.75	20.5
1.70	23.75	18.75

2. Lab equipment: Ultra-Clear Centrifugation tubes (Beckman Coulter, Inc.; Fullerton, CA; http://www.beckmancoulter.com/), ultracentrifuge (min 140,000 *g*), dialysis cassettes (Pierce, Rockford, IL; http://www.piercenet.com/) and Microcon 3500 MWCO Slide-A-Lyzer® tubes (Millipore Corp.; Billerica, MA; http:// www.millipore.com/).

2.2 SDS-PAGE Analysis

1. Buffers and reagents:
   - SDS PAGE loading buffer with β-mercaptoethanol (standard Laemmli loading dye)
   - Standard reagents buffers for silver and Coomassie blue staining.
2. Lab equipment: heating block (95 °C), Standard 1D-gel electrophoresis unit (e.g., Bio-Rad Laboratories; Hercules, CA; http://www.bio-rad.com/).

2.3 Protein Band Isolation and in Gel Digestion

1. Buffers and reagents:
   - 100 mM NH_4HCO_3: 0.79 g/100 ml Milli Q water (stock solution).
   - 20 mM NH_4HCO_3: 10 ml of 100 mM NH_4HCO_3 (dilute from 100 mM stock solution).
   - 55 mM iodoacetamide (IAA) in 100 mM NH_4HCO_3: 0.01 g/ml of 100 mM NH_4HCO_3 (prepared just prior to use).
   - 10 mM dithiothreitol (DTT) in 100 mM NH_4HCO_3: 0.0015 g DTT/ml in 100 mM NH_4HCO_3 (prepared just prior to use).
   - 50 mM acetic acid.
   - 50 mM NH_4HCO_3: (prepared from 100 mM stock solution. Regularly verify the pH 8).
   - 133 mM NH_4HCO_3: 1.05 g NH_4HCO_3/100 ml Milli Q water.
   - Trypsin Gold (Promega): resuspend 20 μg lyophilized trypsin in 1 ml of 50 mM acetic acid. Aliquot and store at −80 °C.

– Trypsin digestion buffer: 12.5 ng trypsin/μl : 150 μl trypsin 20 μg/ml + 90 μl 133 mM NH_4HCO_3 (prepare just prior to use).
– 5% formic acid in 50% acetonitrile: add 5 ml formic acid and 50 ml acetonitrile to Milli Q water to a total volume of 100 ml.

2. Lab equipment: water bath 56 °C, an oven at 37 °C, and a sonicator bath.

2.4 Digestion of Whole Phage Particles

1. Buffers and reagents:
   – 50 mM NH_4HCO_3
   – Digestion buffer: 12.5 ng trypsin/μl in 50 mM NH_4HCO_3 (prepare just prior to use).
   – Denaturation buffer: 6 M urea, 5 mM DTT, and 50 mM Tris–HCl (pH 8).
   – Blocking solution: 100 mM IAA in 50 mM NH_4HCO_3 (prepare just prior to use).
2. Lab equipment: liquid nitrogen, water bath 56 °C, an oven at 37 °C, and a sonicator bath.

2.5 Mass Spectrometry

1. Buffers and reagents:
   – ESI sample buffer: 100 mM acetic acid (HAc) containing 4 pg/μl cortisone (internal analytical standard).
   – HPLC solutions: 100 mM HAc in water and 100 mM HAc in acetonitrile.
2. Lab equipment: A wide variety of mass spectrometers suitable for protein identification is currently available. Routine identification of gel-separated protein bands can be performed with devices allowing peptide mass fingerprinting (MS) or, preferably, tandem mass spectrometry (MS/MS). For the WSA approach, LC–MS/MS is mandatory. Technical details of a suitable LC-ESI setup are described *(13)*. The mass spectrometer mentioned referred to in this chapter is an LCQ-Classic (Thermo Electron Corporation, Waltham, MA; http://www.thermo.com/).

3 Methods

3.1 Phage Purification and Concentration

The identification of structural phage proteins relies strongly on the availability of purified phage particles. Typical contaminants of phage stock solutions are host outer membrane proteins or lipoproteins present after bacterial lysis. Abundant host proteins can cause severe background signals, and should be removed prior to the MS analysis. Two successive rounds of CsCl gradient centrifugation, followed by a dialysis to remove the remaining CsCl, results an ultrapure phage suspension for optimal MS analysis. In order to get a clear image of the different structural proteins, a

minimum of 10^{10} phages should be loaded on the SDS-PAGE gel (**Section 3.2**). As a consequence, the phage stock should be further concentrated to reach an optimal phage density.

1. Prepare a CsCl gradient in the Beckman tubes by subsequently adding 5.7 ml of the CsCl solutions underneath each other, starting with the lowest density. [N.B. This protocol is developed for the Beckman type XYZ rotor (6 × 38 ml capacity). The volumes can be adjusted appropriately for smaller capacity rotors.] Carefully add 15.2 ml of the phage stock (containing 0.5 *g* CsCl/ml to avoid an osmotic shock) on top of the 1.30 g/ml solution, avoiding the disturbance of the gradient.
2. Centrifuge the tubes at 140,000 *g* for 3 h at 4 °C.
3. Collect the opalescent phage band (± 2 ml) by carefully removing all the upper layers, and placing the pipette tip just underneath the target band. Alternatively, the desired band can be removed by puncturing the side of the centrifuge tube just below the phage with a 20 gauge needle and syringe.
4. Dilute the phages to a final volume of 15.2 ml and repeat the procedures (1–3).
5. Dialyse of purified phage suspension for three times for 30 min against 250 volumes of dialysis buffer to remove remaining CsCl.
6. Concentrate the solution 10-fold by reducing the volume of the phage stock, e.g., using a Microcon ultrafiltration device (Millipore) or by vacuum centrifugation.
7. Determine the titer of the resulting phage stock by a standard double-layer method.

3.2 SDS-PAGE Separation of Virion Proteins

Structural proteins from distinct gel bands instead of the whole phage proteins allows in depth analysis, since individual protein samples are less complex. Samples with single proteins are suited for rapid PMF identification, e.g., by Maldi-MS. Although more time-consuming, LC–MS/MS is more successful in identification as it generates more spectra, and the spectra also provide peptide sequence data. Consequently, fewer peptide hits are needed for confident identification, and identification of co-migrating lower abundant proteins is possible. Peptide ionization for MS/MS can be performed by MALDI or ESI. Peptide LC for MALDI-MS/MS requires an additional HPLC device and fraction MALDI target spotter, whereas LC usually is an integrated component of ESI-based mass spectrometer systems. We here restrict to the latter approach making use of an LCQ-Classic mass spectrometer.

Aliquots of 15–50 µl phage particles (minimal 10^{12} pfu/ml) are suspended in SDS loading buffer and denatured by 5 min of heat treatment on 95 °C after adding 50 mM β-mercaptoethanol. The proteins of the destabilized phages are then separated on a discontinuous 12% SDS-PAGE gel. Even for large and complex

phages such as EL and ϕKZ, 1D gels of 5–10 cm length have sufficient resolving power for subsequent LC–MS/MS analysis of the structural proteins. Care should be taken that the electrophoresis front (potentially containing small phage proteins) does not run off the gel.

Suitable staining for the SDS-PAGE gels is achieved with a specific MS-compatible silver-staining protocol *(14)* or standard Coomassie staining. Commercial Coomassie stains like Simply Blue SafeStain (Invitrogen Corp., Carlsbad, CA; http://www.invitrogen.com/) are also compatible with further MS analyses.

3.3 Protein Band Isolation and in Gel Digestion

To isolate protein bands from protein gels, a number of commercial alternatives are available like the Onetouch 2D gel spot picker (The Gel Co., San Francisco, CA; http://www.gelcompany.com/). Gel plugs can also simply be taken from a gel using a 1,000 µl micropipette after widening the opening of the tip with a sharp knife. However, most comprehensive results are obtained by slicing the whole lane including the electrophoresis front, into small segments (20–40 in total) using a clean, sterile scalpel. These slices are ∼ 1 mm broad, or up to 3 mm in interband regions. Proteins are digested within the gel by trypsin (Trypsin Gold, Promega Corp., Madison, WI; http://www.promega.com), which cleaves specifically C-terminal to arginine and lysine residues. Although trypsin digestion is sufficient for a reliable identification of phage proteins, the use of other digesting enzymes (e.g. AspN) with different specificity may result in the identification of additional peptides, resulting in more complete sequence coverage.

Every liquid in the protocol below should be added in such an amount that the gel slices are completely submerged. Gel slices between one and three mm require about 20–50 µl of buffer, respectively, thus avoiding excess quantities of buffer. All gel manipulations are preferentially performed in a laminar flow cabinet so as to avoid contamination of the samples by keratins.

1. Submerge the slice (between 20 and 50 µl/slice) in NH_4HCO_3 in 50% acetonitrile to the gel slices in numbered 1.5 ml Eppendorf tubes, and incubate 10 min at room temperature.
2. Discard the liquid, and repeat step 1 until all the Coomassie blue is completely removed from the gel slices.
3. Dry the slices in a vacuum centrifuge.
4. Submerge the slice (between 20 and 50 µl/slice) in 10 mM DTT in 100 mM NH_4HCO_3 to reduce all disulfide bounds.
5. Incubate for 1 h at 56 °C and subsequently cool to room temperature.
6. Discard the liquid, and submerge the slice (between 20 and 50 µl/slice) in 55 mM iodoacetamide in 100 mM

NH_4HCO_3 to covalently modify cysteine residues (to S-carboxymethyl cysteine) and prevent reformation of disulfide bonds.
7. Incubate 45 min in the dark, and mix every 10 min by briefly vortexing.
8. Discard the liquid.
9. Add 100 µl 100 mM NH_4HCO_3, incubate 10 min and remove the liquid (washing).
10. Add 100 µl acetonitrile, incubate 10 min and discard the liquid (dehydration).
11. Repeat step 9 and 10
12. Dry using vacuum centrifugation.
13. Submerge in digestion buffer and incubate 45 min on ice, so that the gel slices can rehydrate and adsorb the trypsin.
14. Submerge the slice (between 20 and 50 µl/slice) in 50 mM NH_4HCO_3 and incubate overnight at 37 °C.
15. Collect and save the supernatant containing the tryptic peptides in a new Eppendorf tube, one for each gel slice.
16. Submerge the slice (between 20 and 50 µl/slice) in 20 mM NH_4HCO_3, sonicate for 20 min and collect the supernatant in the appropriate tube.
17. Submerge the slice (between 20 and 50 µl/slice) in 5% formic acid in 50% acetonitrile, sonicate for 20 min and collect the supernatant in the appropriate tube.
18. Repeat step 17
19. Save the supernatant at −20 °C until MS analysis is performed.

3.4 Digestion of Whole Phage Particles (WSA Approach)

An alternative/complementary approach for the identification of phage proteins is the digestion of whole phage particles, instead of analysing individual proteins separated on an SDS-PAGE gel. Although whole phages generate a more complex peptide mixture, it is proven that nearly all predicted structural phage proteins can be identified, albeit with lower sequence coverage compared to the SDS-PAGE approach *(10)*. An additional advantage is the need of only one major digest, implying a simplified sample preparation. A more elaborate peptide separation is directly performed on the complex sample by combined reversed-phase HPLC and gas phase fractionation in the mass spectrometer prior to MS/MS analysis.

1. Add 25 µl of digestion buffer to 1–10 µl of phages (at least 10^{10} pfu)
2. Destabilize the particles by 5 successive rounds of freeze-thawing in fluid nitrogen and a 37 °C oven, respectively.
3. Incubate 1 h on 60 °C for complete reduction of the phages.
4. Add 25 µl blocking solution and 150 µl 50 mM NH_4HCO_3.
5. Incubate 45 min on room temperature in the dark, mix every 10 min.

6. Add 40 µl trypsin (20 µg/ml).
7. Incubate overnight at 37 °C.
8. Store at −18 °C until MS analysis is performed.

3.5 Mass Spectrometry

Digested protein samples can be analyzed by MS with ESI peptide ionization [for a review on these techniques see *(15)*]. In a first step, the samples generated in **Sections 3.3** and **3.4** are dried by vacuum centrifugation.

1. Reconstitute the samples in 20 µl 100 mM HAc buffer containing 4 pg/µl cortisone (internal analytical standard).
2. The peptides in the sample are separated by reverse-phase HLPC over a C18 analytical column, using a linear gradient from 5 to 60% (v/v) acetonitrile in water containing 100 mM acetic acid and a run time of 60 min (simple protein samples) or 2 h (WSA samples).
3. The eluate is directly electrosprayed into the mass spectrometer, which is operated in a data-dependent acquisition mode to automatically switch between MS (*m/z* 300–1,500 Thompson in centroid mode at a maximum injection time of 150 ms for standard protein digests) and MS/MS acquisition on the three most intense precursor ions. In analysing the complex peptides mixtures of the WSA samples, an extra gas phase mass-fractionation of the peptides in the LCQ ion trap is performed. Instead of the full scan mass range in the standard acquisition method (350–1,500 Da); the mass range is restricted to one of six specific mass windows (400–600, 600–700, 700–800, 800–900, 900–1,020 and 1,020–1,400 Da). Therefore, the sample is diluted sevenfold in 100 mM HAc containing internal standard. Ten microliters of sample aliquots are analyzed in each of the six mass windows and one aliquot is analyzed in the standard full mass range of the LCQ.

3.6 Protein Identification

ESI–MS/MS spectra are routinely analyzed using Sequest (Thermo Electron Corp.) and Mascot (Matrix Science, Inc., Boston, MA; http://www.matrixscience.com/) search engines against an appropriately customized protein database containing all GenBank bacteriophage sequences as well as those of the bacterial host species. Eventually, the database is supplemented with predicted ORFs from unpublished newly sequenced phage genomes.

Considering the Sequest parameters for LCQ-generated spectra, the cross-correlation value (X_{corr}) is set at ≥ 1.8, ≥ 2.5 or ≥ 3.5 for singly, doubly or triply charged ions, respectively. The delta correlation value (C_{n}) is > 0.1, and parent and fragment ion mass tolerance are 3 and 1 Da, respectively. Possible chemical modifications that are routinely included in the analysis are cysteine carbamidomethylation and oxidation of methionine, histidine, and tryptophan.

For Mascot search, the significance threshold is set at $P \leq 0.05$, parent and peptide ion mass tolerance are ± 3 and ± 0.5 Da, respectively, and one missed trypsin cleavage is allowed.

Especially WSA samples commonly yield single- and double-peptide protein identifications *(10)*. To validate the identification of these proteins, the corresponding spectra can be re-examined with a de novo-sequencing algorithm, e.g., Lutefisk1900 v.1.3.2 *(16)* utilizing the database sequence option. In doing so, the de novo-derived sequence candidates are evaluated against the peptide sequence (as returned by Sequest and Mascot) entered in the Lutefisk database file. In case the program evaluates the database sequence as being good as or better than the de novo sequences, the corresponding single- and double-peptide protein identification can be assumed to be valid.

References

1. 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.
2. Loessner, M.J., Inman, R.B., Lauer, P. and Calendar, R. (2000) Complete nucleotide sequence, molecular analysis and genome structure of bacteriophage A118 of *Listeria monocytogenes*: implications for phage evolution. *Mol Microbiol.*, **35**, 324–340.
3. Zimmer, M., Sattelberger, E., Inman, R.B., Calendar, R. and Loessner, M.J. (2003) Genome and proteome of *Listeria monocytogenes* phage PSA: an unusual case for programmed + 1 translational frameshifting in structural protein synthesis. *Mol Microbiol.*, **50**, 303–317.
4. Roberts, M.D., Martin, N.L. and Kropinski, A.M. (2004) The genome and proteome of coliphage T1. *Virology*, **318**, 245–266.
5. Chibani-Chennoufi, S., Canchaya, C., Bruttin, A. and Brüssow, H. (2004) Comparative genomics of the T4-Like *Escherichia coli* phage JS98: implications for the evolution of T4 phages. *J Bacteriol.*, **186**, 8276–8286.
6. Scholl, D., Kieleczawa, J., Kemp, P., Rush, J., Richardson, C.C., Merril, C., Adhya, S. and Molineux, I.J. (2004) Genomic analysis of bacteriophages SP6 and K1–5, an estranged subgroup of the T7 supergroup. *J Mol Biol.*, **335**, 1151–1171.
7. Levesque, C., Duplessis, M., Labonte, J., Labrie, S., Fremaux, C., Tremblay, D. and Moineau, S. (2005) Genomic organization and molecular analysis of virulent bacteriophage 2972 infecting an exopolysaccharide-producing *Streptococcus thermophilus* strain. *Appl Environ Microbiol.*, **71**, 4057–4068.
8. Byrne, M. and Kropinski, A.M. (2005) The genome of the *Pseudomonas aeruginosa* generalized transducing bacteriophage F116. *Gene*, **346**, 187–194.
9. Bukovska, G., Klucar, L., Vlcek, C., Adamovic, J., Turna, J. and Timko, J. (2006) Complete nucleotide sequence and genome analysis of bacteriophage BFK20–a lytic phage of the industrial producer *Brevibacterium flavum*. *Virology*, **348**, 57–71.
10. Lavigne, R., Noben, J.P., Hertveldt, K., Ceyssens, P-J., Briers, Y., Dumont, D., Roucourt, B., Krylov, V.N., Mesyanzhinov, V.V., Robben, J. and Volckaert, G. (2006) The structural proteome of *Pseudomonas aeruginosa* bacteriophage ϕKMV. *Microbiology*, **152(Pt 2)**, 529–534.
11. Ceyssens, P-J., Lavigne, R., Chibeu, A., Mattheus, W., Hertveldt, K., Robben, J. en Volckaert, G (2006) Genomic analysis of *Pseudomonas aeruginosa* phages LKD16 and LKA1: Establishment of the ϕKMV subgroup within the T7 supergroup. *J. Bacteriology*, **188(19)**, 6924–6931.
12. Rees, J.C. and Voorhees, K.J. (2005) Simultaneous detection of two bacterial pathogens using bacteriophage amplification coupled with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.*, **19**, 2757—2761.

13. Dumont, D., Noben, J.P., Raus, J., Stinissen, P. and Robben, J. (2004) Proteomic analysis of cerebrospinal fluid from multiple sclerosis patients. *Proteomics*, **4**, 2117–2124.
14. Shevchenko, A., Wilm, M., Vorm, O. and Mann M. (1996) Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal Chem.*, **68**, 850–858.
15. Steen, H. and Mann, M.(2004) The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol.*, **5**, 699–711.
16. Taylor, J.A. and Johnson, R.S. (1997) Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom.*, **11**, 1067–1075.

Section III

Community Bacteriophage Approaches

Chapter 15

Isolation Independent Methods of Characterizing Phage Communities 1: Strain Typing Using Fingerprinting Methods

Clemens Pausz, Jessica L. Clasen and Curtis A. Suttle

Abstract

Since most of the phage genomes isolated from natural samples are previously unknown sequences, an isolation-independent approach is necessary to quantify the diversity of natural viral communities. Currently, two different methodological approaches are widely used to obtain genetic fingerprints of natural phage communities. While the separation of different viral genomes with pulsed field gel electrophoresis (PFGE) is based on the size of the genome, denaturing gradient gel electrophoresis (DGGE) uses minor differences in gene base composition to separate fragments of amplified DNA from natural viral communities. Finger printing techniques are a relatively fast and cheap tool to assess the diversity of environmental viruses. Together, PFGE and DGGE provide useful tools to study viral ecology in natural habitats.

Key words: Viruses, bacteriophages, virioplankton, phage community composition, genetic fingerprints, DGGE, PFGE.

1 Introduction

Viruses are the most abundant biological entity on the planet *(1–3)*. For example, in terrestrial ecosystems, the numbers of virus-like particles range between $8.7 \times 10^8 - 1.1 \times 10^9\ g^{-1}$ dry weight in agricultural soils and $3.1 \times 10^9 - 4.17 \times 10^9\ g^{-1}$ dry weight in forest soils *(4)*. Current estimates of viral abundance in aquatic systems range from 3×10^6 to 1×10^8 viral particles mL^{-1} with bacteriophages being the principal component of virioplankton communities *(4–7)*. These phages are an important mortality factor for prokaryotes *(3, 8)*, affecting the abundance and diversity of microbial communities as well as the cycling of carbon and nutrients *(9–12)*.

Martha R. J. Clokie, Andrew M. Kropinski (eds.), *Bacteriophages: Methods and Protocols, Volume 2: Molecular and Applied Aspects, vol. 502*, © 2009 Humana Press, a part of Springer Science+Business Media
DOI 10.1007/978-1-60327-565-1_15 Springerprotocols.com

As a result of the narrow host range of bacteriophages *(13–15)* and algal viruses *(16–18)*, a natural viral community is at least as diverse as the corresponding host community. Therefore, natural viral communities show very high genetic richness. For example, metagenomic studies of coastal waters and sediments showed that 60–80% of viral sequences were not significantly similar to other sequences in public databases *(19)*. Consequently, isolation-independent methods are an invaluable tool for assessing the composition and diversity of phage communities.

1.1 Separation of Different-Sized Viral Genomes by Pulsed-field Gel Electrophoresis (PFGE)

PFGE enables fractionation of intact large DNAs based on genome size. This method has been used to separate and genetically fingerprint viral genomes in natural samples *(5, 20)* and has been described in detail elsewhere *(21)*. Viral genomes vary significantly in size, therefore providing a basis for the separation of different genomes by gel-electrophoresis. The strength of this method lies in the separation of intact genomes and its independence from a PCR-step requiring a homologous gene. The major disadvantage is the relatively low sensitivity of the method, requiring the viral community to be concentrated.

Viruses from natural samples are first concentrated and then immobilized in agarose blocks. The DNA is extracted in the agarose blocks, which are then inserted into the wells of an agarose gel and PFGE is performed to separate the intact viral genomes.

The major steps of this method are (1) collecting and concentrating viral communities from natural samples, (2) extracting DNA from the viral particles, (3) PFGE, and (4) documentation and interpretation of the gels.

1.2 Separation of PCR Amplified Gene Fragments by Denaturing Gradient Gel Electrophoresis (DGGE)

DGGE separates dsDNA fragments based on sequence. DNA denatures as it passes through a polyacrylamide gel that has an increasing gradient of urea and formamide, causing the sequence fragments to be separated based on nucleotide composition. DGGE is extremely sensitive, having the potential to separate DNA fragments that differ by a single base pair *(22)*, and providing an approach to fingerprint viral communities *(23–25)*. The following section describes a procedure for using DGGE to fingerprint a subset of natural viral communities. The primers used in this case (CPS-4 and CPS-9) are degenerate (1,024-fold degeneracy) and were designed to amplify a 595 bp region of the capsid assembly protein gp20 of cyanomyoviruses, but they likely target a broader suite of myoviruses *(26)*. For other primers see **Table 15.5 and 15.6**. The PCR method listed below is simplified and assumes that extracted environmental DNA will be used as a template.

The individual steps of DGGE are described in detail below, and include assembling and casting the gel, preparing the

electrophoresis tank, and loading, running, staining, destaining, and viewing the gel.

2 Materials for Pulsed-Field Gel Electrophoresis *(PFGE)*

2.1 Purification and Concentration of Viral Particles

2.1.1 Water Sample

1. Two tripod stainless steel disc filter holders, filter-diameter 142 mm (e.g., Millipore, Billerica, MA or Pall Corp., East Hills, NY).
2. Vacuum pump.
3. Tangential flow ultrafiltration system with a 30,000 kDa cut-off cartridge and corresponding cartridge holder or cartridge header set (e.g., Prep/Scale TF/F or Helicon S10 or S40—both Millipore).
4. Peristaltic pump (e.g., Masterflex I/P Easyload pump head + Standard BDC Drive—Cole Parmer, Vernon Hills, IL).
5. Appropriate tubing: e.g., Norprene™ Pharmed™ for use in pump head tubing retainer (provided by pump manufacturer), Tygon™ B44-3 with ~0.094 in. wall thickness for the feeding (available from Cole-Parmer).
6. GF/C glass-fiber filters, 142 mm diameter (Whatman, Brentford, UK).
7. 0.2 μm pore-sized polyvinyldifluoride Durapore® filters, 142 mm diameter (Millipore).
8. Centrifuge equipped with a swinging bucket rotor for conical 50 mL tubes.
9. Large volume (15 mL) centrifugal filter devices (e.g., Amicon® Ultra-15 30 kDa cut-off–Millipore).

2.1.2 Soil Sample

1. 25 mL Teflon-coated polyethylene centrifuge tubes.
2. Eluent: 1% potassium citrate (10 g potassium citrate, 1.44 g $Na_2HPO_4.7H_2O$, 0.24 g KH_2PO_4, L^{-1} (pH 7) or 10 mM sodium pyrophosphate (pH 7), or 250 mM glycine (pH 8).
3. Vortex (e.g., Fisher Vortex Genie 2, Fisher Scientific, Hampton, NH).
4. Sonicator (e.g., Branson 185 Sonifier–Branson Ultrasonics, Danbury, CT).
5. Microcentrifuge.

2.2 Preparation of Viral Particles for PFGE

1. Low melting point agarose (e.g., In-Cert agarose, FMC Bioproducts, Rockland, ME).
2. Plug moulds (BioRAD, provided with PFGE-system).
3. Digestion buffer (250 mM EDTA [pH 8.0], 1% sodium dodecyl sulfate, and 1 mg mL^{-1} proteinase K; Invitrogen, Carlsbad, CA).

4. SM buffer (0.1 M NaCl, 8 mM $MgSO_4 \cdot 7H_2O$, 50 mM Tris–HCl [pH7.5] and 0.005% (wt/vol) glycerol).
5. TE buffer (10 mM Tris–Cl [pH 8.0], 1 mM EDTA).
6. Washing buffer (10 mM Tris–HCl and 1 mM EDTA [pH 8.0]).
7. Storage buffer (20 mM Tris–HCl and 50 mM EDTA [pH 8.0]; can be replaced by washing buffer).

2.3 PFGE

1. Pulsed field gel electrophoresis system (e.g., DR-II Cell, BioRAD, Hercules, CA) including electrophoresis chamber, pump, cooling unit, control module;
2. PFGE-grade Agarose (e.g., SeaKem GTG—Cambrex, Rockland, ME).
3. Gel buffer (1 × TBE: 90 mM Tris–borate and 1 mM EDTA [pH 8.0])
4. 10 × x running buffer solution (10 × TBE: contains 108 g Tris–HCl, 55 g H_3BO_3, 40 mL of 0.5 M EDTA [pH 8.0]).
5. 10× loading buffer stock solution (25% Ficoll and 0.25% xylene cyanol).
6. Standards: MidRange PFG Marker I (15–300 kb), λ-DNA (both New England Biolabs, Ipswich, MA), 5 kb ladders (Bio-Rad, Hercules, CA).

2.4 Documentation and Interpretation of PFGE-Gels

1. Gel-documentation system (e.g., AlphaImager^{tm} 3400. Alpha Innotech, San Leandro, CA).
2. Nucleic acid staining solution (e.g., ethidium bromide: 5 mg mL^{-1}).

3 Materials for the Separation of PCR Amplified Gene Fragments by Denaturing Gradient Gel Electrophoresis (DGGE)

1. Use materials described in Section 2.1.

3.1 Harvesting and Concentration of Viral Particles

1. Microcentrifuge tubes (2 mL, with o-ring caps).
2. Table centrifuge with rotor for microcentrifuge tubes.
3. Water bath or heating block for microcentrifuge tubes
4. TE buffer (10 mM Tris–Cl [pH 7.4] and 1 mM EDTA).
5. Sodium dodecyl sulfate (SDS) (10% wt/vol).
6. 0.5 M EDTA.
7. Proteinase K (20 mg/mL; keep frozen at −20 °C; Invitrogen, Carlsbad, CA).
8. 5 M NaCl.
9. CTAB/NaCl solution (10% CTAB in 0.7 M NaCl).

3.2 Extraction of viral DNA

10. Buffer (Tris/HCl, pH 8.0) saturated phenol (e.g., Invitrogen).
11. PCI (25:24:1 buffer-saturated phenol/chloroform/isoamyl alcohol).

12. CI (24:1 chloroform/isoamyl alcohol).
13. 3 M Sodium acetate.
14. Ethanol (100%, room temperature).
15. Ethanol (70%, −20 °C).

3.3 Polymerase Chain Reaction (PCR) with Cyanophage-specific Primers (CPS)

1. 0.5 mL sterile microfuge tubes.
2. 20–200 μL pipette and sterile tips.
3. 0. 5–5 μL pipette and sterile tips.
4. PCR Thermocycler (PCR Express, Hybaid Limited, Middlesex, UK).
5. Electrophoresis gel box and associated equipment (combs, etc.).
6. DC power supply.
7. UV transilluminator.
8. Gel doc system (e.g., AlphaImager^{tm} 3400, Alpha Innotech, San Leandro, CA).
9. Glass Pasteur pipettes.
10. PLATINUM *Taq* polymerase (1 unit per reaction) (Invitrogen Corporation, Carlsbad, CA).
11. 10 × PCR Buffer (without $MgCl_2$) (comes with *Taq*).
12. 50 mM $MgCl_2$ (comes with *Taq*).
13. 2 mM dNTPs: Make a stock solution of 20 mM dNTPs by mixing 100 μL of each base (dATP, dGTP, dCTP, and dTTP; available from Invitrogen Corporation, Carlsbad, CA) with 100 μL of Milli-Q (**Note 1**). Make a 1:10 dilution of this stock solution (final concentration of working solution = 2 mM), aliquot, and store at −20 °C.
14. 10 μM CPS-4 *(27)* (5′-CATWTCWTCCCAHTCTTC-3′): Make a 100 μM stock solution by dissolving the primer with Milli-Q water (**Note 1**). Dilute 1:10 to make a working solution.
15. 10 μM CPS-9 *(1)* (5′-SWRAAATAYTTICCRACRWAGGATC-3′): Make a 100 μM stock solution by dissolving the primer in Milli-Q water (**Note 1**). Dilute 1:10 to make a working solution.
16. Milli-Q water (*see* **Note 1**).
17. 0.5× TBE from 10× TBE: To make 10× TBE mix 108 g of tris base, 55 g of boric acid, 40 mL of 0.5 M EDTA (pH 8) and bring to 1 L with Milli-Q. Add 50 mL of 10× TBE to 950 mL of Milli-Q to make 1 L of 0.5× TBE.
18. 1.5% Agarose gel: 1.5 g of agarose (Invitrogen Corporation, Carlsbad, CA) into 100 mL of 0.5×TBE. Melt agarose in microwave (∼2 min at high). Wait until cool to touch and then pour the gel.
19. Loading buffer: To make a 6× stock solution of loading buffer mix 0.4% w/v bromophenol blue in 30% glycerol. Dilute to 1× working solution.

20. Ethidium bromide solution: 25 μL of stock EtBr solution (10 mg/mL) to 500 mL of Milli-Q. Replace as needed. (EtBr is a carcinogenic chemical, wear gloves, and labcoat at all times when handling.)

3.4 Denaturing Gradient Gel Electrophoresis (DGGE)

3.4.1 Assembling and Casting a DGGE Gel

1. 0.45 μm pore-size polycarbonate filter.
2. Two front glass plates (10 cm)*.
3. Two back glass plates (16 cm)*.
4. Four spacers*.
5. Four sandwich clamps*.
6. Casting stand*.
7. Alignment card or a thick piece of paper*.
8. Rubber mat*.
9. Tygon tubing (three pieces: one 9 cm piece and two 15.5 cm pieces)*.
10. 19 gauge needle.
11. Tape.
12. Three-way port or Y-fitting*.
13. Gradient delivery system*.
14. 2 × 60 cc Falcon^{tm} tubes.
15. 2 × 30 cc Syringes.
16. Two syringe sleeves*.
17. Two plunger caps*.
18. A beaker.
19. Well combs*.
20. 50× TAE: Mix 242 g of Tris base, 57.1 mL of glacial acetic acid, 37.2 g of $Na_2EDTA{\cdot}2H_2O$, and bring volume up to 1 L with Milli-Q. Adjust the pH to 8.5.
21. Denaturing solutions: A 100% denaturing solution contains 7 M urea and 40% deionized formamide. With CPS primers, the high denaturing solution (HDS) is a 40% denaturing solution with 8% polyacrylmide (**Note 2**), while the low denaturing solution (LDS) is a 20% denaturing solution with 7% polyacrylmide. To make 100 mL of HDS, add 20 mL of 40% acrylamide/bis (37.5:1, 2.6% C), 2 mL of 50× TAE, 16 mL of deionized formamide, and 16.8 g of urea to ~75 mL of Milli-Q. The procedure is the same to make 100 mL of LDS except the acrylamide/bis, deionized formamide and urea are reduced to 17.5 mL, 8 mL and 8.4 g, respectively. Bring the volumes up to 100 mL and make sure the chemicals are completely dissolved. Filter the solutions through 0.45 μm pore size nitrocellulose filters and store at 4 °C in dark bottles. The denaturing solutions are good for 1 month. Acrylamide and formamide are hazardous chemicals. Work in a fume hood, and wear protective goggles, gloves, and a labcoat.
22. 70% Isopropanol: Mix 70 mL of iso-amyl alcohol with 30 mL of Milli-Q. This solution is flammable.

23. 10% w/v Ammonium persulfate: Add 0.03 g of ammonium persulfate to 300 μL of Milli-Q water. Dissolve completely. Make this solution fresh daily.
24. TEMED: *N*, *N*, *N*′-tetra-methyl-ethylenediamine.

3.4.2 Preparing and Pre-Warming the Tank Buffer

1. Electrophoresis core*.
2. Electrophoresis tank*.
3. Heater/pump unit*.
4. Lid stand*.
5. 1× TAE from 50× TAE: For every liter, add 20 mL of 50× TAE (*see* **Section 2.2.1** above) to 980 mL of Milli-Q water. Approximately 8 L are needed for each gel.

3.4.3 Loading and Running the DGGE Gel

1. Flat-tipped loading pipette tips, 0.4 mm thick, 200 μL flat gel tips, Electrophoresis DC power supply.
2. Loading buffer: See **Section 2.1**. Use a final concentration of 1×.

3.4.4 Staining the DGGE Gel

1. Plastic container with lid for stain.
2. Plastic container with lid for destain.
3. Gel doc system (e.g., AlphaImager^{tm} 3400. Alpha Innotech, San Leandro, CA).
4. Transilluminator with filter for SYBR Green (excitation 497 nm and emission 520 nm).
5. SYBR Green staining solution: Add 4 μL of 10,000× SYBR Green I (Invitrogen) to 500 mL of 1× TAE. SYBR Green is light sensitive; work in dim light. The 1× TAE must be newly spiked with SYBR Green for each DGGE gel, but replace the 1× TAE when needed. SYBR Green is a nucleic acid stain use with caution.
6. 1× TAE from 50× TAE: See **Section 3.4.2**.

4 Methods for Pulsed-Field Gel Electrophoresis (PFGE)

4.1 Purification and Concentration of Viral Communities

4.1.1 Concentration of Natural Viral Communities from a Water Sample

Viral particles are concentrated using a two-stage ultrafiltration system *(28)* as previously described *(29)*. The volume of water sampled depends on the required amount of viral DNA and the efficiency of the concentration process (**Note 1**).

1. Collect a water sample of 20–200 L, the volume depends on the *in situ* abundance of viruses. The smaller volume would be typical for a coastal water sample, while 200 L would be suitable for an open-ocean sample.
2. Prefilter the sample through two 142 mm diameter stainless-steel filter holders in series containing 1.2 μm nominal pore-size GF/C glass-fiber and 0.2 μm pore-size

polyvinyldifluoride membrane filters to remove zooplankton, phytoplankton (**Note 2**), and bacterioplankton, respectively. Collect the filtrate in an acid-rinsed polycarbonate container with at least 20 L volume. (Since ultrafiltration and prefiltration are run simultaneously, a 20 L container is sufficient.)

3. Once the prefiltration is started, set up the ultrafiltration system. The remaining particulate matter in the 0.2 μm filtrate is concentrated using a 30,000 kDa cut-off ultrafiltration system (e.g., Prep/Scale-TF/F with cartridge holder—Millipore).
4. Use the peristaltic pump to push the 0.2 μm filtrate from the container through the ultrafiltration system (initial flow rate ~1.5–2 L min^{-1}; backpressure should not exceed 20 psi). The ultrafiltrate can be disposed of or kept (e.g., for seawater media). The retentate is returned to the container with the filtrate. Once the retentate reaches 3 L, transfer the volume to a large plastic beaker, and continue ultrafiltration until the retentate reaches ~250–500 mL. Flush the cartridge with a small volume of ultrafiltrate. Stop the ultrafiltration and drain the hold-up volume from the cartridge and tubing into the beaker.

To determine the concentration efficiency take 1–5 mL samples of the original water sample, filtrate, concentrate, and ultrafiltrate and measure viral abundances by epifluorescence microscopy following established protocols as outlined elsewhere in this book *(30–32)*.

Record the volumes of sample, prefiltrate, and retentate in order to calculate the overall concentration efficiency. The final viral abundance is required to calculate the amount of sample necessary for PFGE.

4.1.2 Concentration of Natural Viral Communities from a Soil Sample

This protocol closely follows *(4)*

1. Weigh 5 g-samples of moist soil into 25 mL Teflon-coated polyethylene centrifuge tubes.
2. Add 15 mL of eluent; either 1% potassium citrate (10 g potassium citrate, 1.44 g $Na_2HPO_4.7H_2O$, and 0.24 g KH_2PO_4, L^{-1}, pH 7), 10 mM sodium pyrophosphate (pH 7), or 250 mM glycine (pH 8).
3. Vortex tubes.
4. Sonicate tubes on ice for 3 min (gently shake tube for 30 s each minute).
5. Spin down soil particles by centrifuging at $10,000 \times g$.
6. Pass supernatant through 0.2 μm pore-size syringe filters to remove bacteria and small soil particles and collect in a 50 mL tube.

Resuspend soil pellets in fresh eluent and repeat steps 2 to 6 twice. Pool supernatants in a 50 mL tube.

4.2 Preparation of Viral Concentrates for PFGE

This protocol follows *(5)*, *(20)*, *(21)*

4.2.1 Preparation of Viral Concentrates for PFGE as Embedded Samples (Note 3)

1. The viruses in ~15 mL of retentate (depending on the virus abundance) are further concentrated in a 50 mL centrifugal filter by centrifugation for 12 min at $3,220 \times g$. The filtrate is gently decanted, and an additional 15 mL of sample is added to the tube. Repeat two or more times depending on the abundance of viruses in the concentrate (~10^9 total virus particles are typically required).
2. Finally, add 15 mL of SM buffer and centrifuge for 12 min at $3,220 \times g$. Repeat.
3. The concentrate is transferred to a 1.5 mL screw cap centrifuge tube using a pipettor and disposable pipette tips, and stored at 4 °C.
4. Assess viral abundance in the concentrate.
5. Take an aliquot of the concentrate containing ~10^9 viral particles; adjust the volume to 50 µL, and mix it with 50 µL of molten (50 °C) 1.5% low melting point agarose; gently vortex for 5 s or mix by inverting (**Note 4**).
6. Dispense the mixture into plug molds using a pipettor with 200 µL pipette tips.
7. When the gel is solidified (~1 h), punch the plugs from the molds into ~200 µL of digestion buffer, and incubate in the dark at room temperature overnight.
8. Decant the digestion buffer and wash plugs in washing buffer. Repeat three times.
9. At this point, the agarose plugs can be stored at 4 °C in 1–2 mL of storage buffer.

4.2.2 Preparation of Viral Concentrates for PFGE as Liquid Samples

Follow steps 1–4 in **Section 3.1.2** but instead of SM buffer use TE buffer.

5. Take an aliquot of the concentrate containing ~10^9 viral particles; adjust the volume to 50 µL and heat the sample in a water bath at 60 °C for 10 min.
6. Centrifuge sample for 30 s to spin down potential condensation and transfer into a clean tube.
7. Cool sample on ice.

4.3 PFGE

These instructions are based on the BDR-II cell pulsed field gel electrophoresis system.

1. Prepare a 5-mm thick 1% agarose gel using a 1% agarose solution prepared using PFGE-grade agarose and 1× TBE. Assemble the casting stand provided with the PFGE system and position it with the provided level. Fill the stand with the agarose solution up to a level of about 5 mm (~92–100 mL of liquid gel).
2. Insert the comb and let the gel solidify for 1–2 h.

3. While gel solidifies, fill the electric-field cell with $0.5\times$ TBE running buffer. Switch on the pump and set the flow rate to 80–90% of the maximum ($\sim$1 L min^{-1}). Adjust the temperature at the cooling unit to 14 °C (**Note 6**).
4. Remove the comb from the solidified gel and fill the wells with running buffer using disposable Pasteur pipettes. Make sure there are no air bubbles in the wells.
5. Carefully remove the sample plugs from the storage tubes, dry them with a Kimwipe^{TM} (**Note 6**), and place them into the wells (liquid samples *see (10)*).
6. A standard consisting of phages with known genome sizes and concentrations is helpful for quantitative analysis of PFGE results. The standard is processed in the same way as the viral samples and loaded onto the PFGE gels.
7. Cut small slices ($\sim$1 mm thick) of λ-DNA and fill wells in the center and outermost lanes of the gel. (Other mid-range markers can be used additionally.)
8. Remove the frame from the casting stand, making sure the gel is attached tightly to the platform (**Note 8**). Put the platform with the attached gel into the PFGE cell.
9. Use loading buffer to add the 5 kb mass ladder in the wells adjacent to the λ-DNA (**Note 7**).
10. Use 5 μL loading buffer to load liquid samples.
11. Adjust the running conditions on the control unit. Choose "contour-clamped homogeneous electric field;" 6 V cm^{-1}; pulses from 1 to 10 s with linear ramp; reorientation angle 120° (**Note 9**).
12. Close the lid of the PFGE-cell and start the run (**Note 10**).
13. After the run remove the platform and the gel from the cell and remove the gel from the platform. Stain the gel in ethidium bromide (5 μg mL^{-1} for 30 min and shortly destain in Milli-Q water).
14. Scan the gel and store the image using a gel documentation system (e.g., AlphaImager^{tm} 3400).

4.4 Documentation and Interpretation of PFGE-Gels and Fingerprints

1. Qualitative and quantitative analysis can be performed on the scanned images.
2. A straight-forward qualitative approach to compare samples is to calculate binary similarity coefficients based on the presence or absence of bands (**Note 11**) *(5)*.
3. For more in-depth quantitative analysis compare the relative fluorescence of each band with standards of known viral concentration and genome size. Subsequently, the concentration of viral particles in single bands is calculated applying the equation: $V_u G_u / A_u = V_k G_k / A_k$ (where V = number of viruses in a band, G = genome size in kb pairs of DNA obtained from the PFGE gel, A = intensity, u = unknown viruses, k = known virus standard) *(5)*.

4. To compare results from different samples and gels, calculate the relative abundance of particles in each band as fraction of the total number of particles in the fingerprint.

5 Methods for the Separation of PCR Amplified Gene Fragments by Denaturing Gradient Gel Electrophoresis (DGGE)

5.1 Harvesting and Concentration of Viral Particles

1. Follow the procedure described in **Section 2.1.1**.

5.2 Extraction of Viral DNA

This protocol closely follows the method described in *(34)*.

1. Transfer 500 μL of viral concentrate to fresh o-ring tubes.
2. Add 25 μL of SDS (10%) and 20 μL of EDTA (0.5 M) and incubate for 15 min at 65 °C.
3. Add 550 mL of phenol, mix by inversion for 1 min.
4. Spin in microcentrifuge at 12,000 rpm for 3 min and transfer 500 μL of the aqueous (top) phase to a fresh tube.
5. Repeat step 5.
6. Add equal volume of PCI (*see* **Section 3.2**) and mix by inversion for 1 min.
7. Spin in microentrifuge at 12,000 rpm for 1 min and transfer 500 μL of the aqueous (top) phase to a fresh tube.
8. Repeat step 7.
9. Add equal volume of CI and mix by inversion for 1 min.
10. Spin in microcentrifuge at 12,000 rpm for 3 min and transfer 500 μL of the aqueous (top) phase to a fresh tube.
11. Add 40 μL NaAc (3 M) and 500 μL of ethanol (100%), mix by inversion.
12. Allow DNA to precipitate at −20 °C overnight.
13. Spin tubes in microcentrifuge at 12,000 rpm and carefully remove ethanol with a pipettor leaving DNA pellets.
14. Add 500 μL of ice-cold ethanol (70%) to dissolve salts for 5 min at room temperature.
15. Spin in microcentrifuge at 12,000 rpm for 3 min and carefully remove ethanol with a pipettor.
16. Allow pellets to dry for 20 min by uncapping tubes and leaving at room temperature.
17. Re-suspend pellets in 100 μL of distilled water or TE-buffer and store at −20 °C.

5.3 Polymerase Chain Reaction with Cyanophage Specific Primers (CPS) 4 and 9

1. For more details on PCR, *see* **Chapter 26**. Recipe for one 50 μL PCR using CPS primers 4 and 9 (**Table 15.1**)

10× PCR buffer	5 μL
50 mM $MgCl_2$	1.5 μL
2 mM dNTPs	5 μL
10 μM CPS-4	1 μL
10 μM CPS-9	1 μL
PLATINUM *Taq* polymerase	0.2 μL

DNA Template (**Note 14**)	2.0 μL
Milli-Q (**Note 12**)	34.3 μL

2. PCR parameters for a PCR with CPS primers 4 and 9.

Initial denaturation	94 °C for 90 s
35 cycles of:	
Denaturation	94 °C for 45 s
Annealing	50 °C for 60 s
Extension	72 °C for 45 s
Final extension	72 °C for 5 min
Hold	4 °C

3. When the PCR is complete, pour a 100 μL 1.5% agarose gel with 0.5× TBE. Load 10 μL of each PCR product and ~2 μL of 6× loading buffer onto the gel. Include a molecular marker such as a 100 bp ladder (Invitrogen, Carlsbad, CA). Run a 10× 10 cm gel at 80 V for 70 min in a 0.5× TBE tank buffer. Stain the gel in EtBr for > 30 min. View stained gel on an UV transilluminator and photograph. Store remaining PCR products at −20 °C until further use.
4. If necessary, a second round PCR can be used to increase yield (**Note 15**). Plug visible bands from the first round PCR with a clean glass pipette. Place the plugged DNA in a sterile microfuge tube and add 100 μL of 0.5× TBE. Heat to 65 °C for 60 min to elute the DNA. Repeat the PCR outlined above but, use 1 μL of eluted DNA as a template in the second round PCR (also increase the amount of Milli-Q added to each reaction to 35.3 μL) and drop the number of PCR cycles to 20. Confirm a clean PCR product by electrophoresis (one clean band at the target base pair size, which in this case is ~600 bp). Store the remaining second round PCR products at −20 °C until ready to be used in a DGGE.

Table 15.1
CPS 4 and 9 denaturing solutions *(37)*, also *see* Note 13

Chemical	20% denaturing and 7% gel	40% denaturing and 8% gel
40% Acrylamide/bis (37.5:1, 2.6% C)	17.5 mL	20 mL
50× TAE	2 mL	2 mL
Formamide	8 mL	16 mL
Urea	8.4 g	16.8 g
dH_2O	To 100 mL	To 100 mL

Filter solutions through 0.45 μm filters and store at 4°.
Solution good for ~1 month.

5.4 Denaturing Gradient Gel Electrophoresis (DGGE) with CPS Amplified PCR Products

5.4.1 Assembling and Casting a DGGE Gel

1. Generally follow the manufacturer's instructions for pouring and assembling the gel, keeping the following suggestions in mind.
2. Clean both the 10 cm front and 16 cm back glass plates with 70% isopropanol and dry thoroughly with Kimwipes™.
3. *Assemble plates*: On a clean surface, lay down the back glass plate. Place the right and left spacers on the outside edges of the back rectangular plate (with the straight edges facing in). Carefully place the front glass plate on top of the spacers and align with the bottom edge. Place the assembled glass plates in the sandwich clamps (with the arrows on the sandwich clamps facing up) and loosely tighten the screws.
4. Place sandwiched glass plates in the casting stand. Use the alignment slot, which is the one without side clamps. Slightly undo the sandwich clamps to allow the glass plate and spacers to self-align along the bottom edge. To ensure that the spacers remain in position, insert an alignment card or piece of thick paper between the glass plates. Tighten sandwich clamps by pushing in at the arrow on the clamps with your palms and tighten the top screws with your thumbs and index fingers. Remove the alignment card or piece of paper. Remove the tightened sandwiched glass plates and run your finger along the bottom to make sure that plates and spacers are flush. If they are not, reassemble plates (a.k.a. repeat steps 1–3).
5. Ensure that the cast stand is level, place the rubber mat in the front slot of the casting stand and place the sandwiched glass plates in the front slot with the short glass plate facing out. Push in the locks and turn 180° to lock the sandwiched glass plates in place on top of the rubber mat.
6. Attach at least 9 cm of Tygon™ tubing to a 19-gauge needle (**Note 16**) and tape the needle to the center of the back glass plate. The beveled side of the needle should point toward you. Attach the Tygon™ tubing to the three-way port and tighten the attachment screw.
7. To pour a 1 mm thick, 16 cm by 16 cm gel set the volume setting indicator located on the gradient delivery system to 14.5 by tightening the volume adjustment screw (**Note 17**).
8. Label a 30 cc syringe "High" and another "Low" for each denaturing solution, and add a syringe sleeve to each. The sleeve should be flush against the top of the syringe and should face away from the volume gradations. Attach a plunger cap to each syringe. The plunger cap should be tight and the back screw (the lever attachment screw) should be parallel to the covered part of the sleeve (**Fig. 15.1**) Ensure

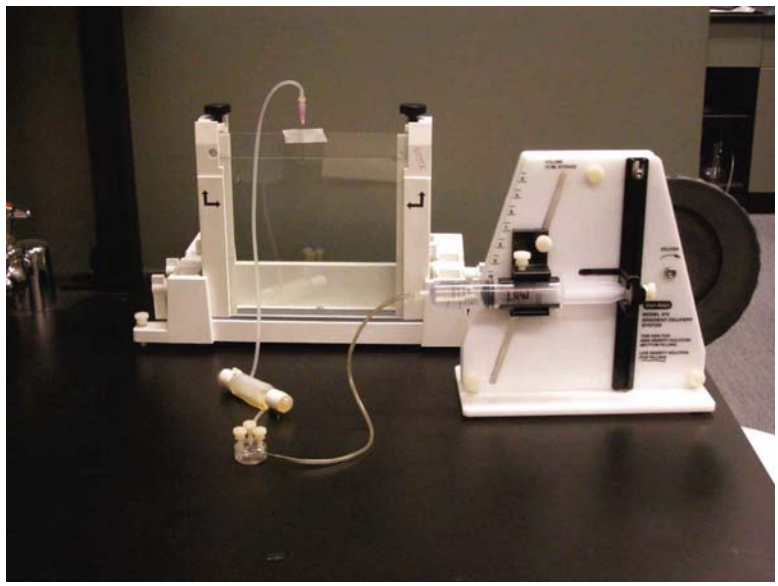


Fig. 15.1. DGGE casting assembly.

that the lever attachment screw can easily fit into the track on the gradient delivery system but do not leave attached to the system. Finally, attach 15.5 cm of Tygon® tubing to each syringe. It is critically important to ensure that all the Tygon™ tubing is clean and dry before proceeding.

9. Cock the wheel (move it in a counter-clockwise direction until it stops) on the gradient delivery system so that it is ready to deliver the denaturing solution.
10. Prepare a beaker of hot water, which will be used immediately following the casting process to clean the syringes and tubes. *Steps 11–16 are time sensitive. Read the directions first before casting the gel. Once the ammonium persulfate and TEMED are added to the denaturing solutions, you have ~7–10 min before the solutions begin to polymerize.*
11. Label two 60 cc Falcon™ tubes "High" and "Low". To each tube, add 16 mL of the corresponding denaturing solution, 100 μL 10% ammonium persulfate and 14 μL of TEMED. Invert carefully four times.
12. Carefully draw the entire "High" solution into the appropriate syringe to avoid creating air bubbles, If necessary, remove air bubbles by inverting the syringe, gently tapping it against a table, and pushing on the syringe until all the air is expelled. Keep the loss of denaturing solution to a minimum.
13. Place the full "High" syringe on the back on the gradient delivery system. Ensure that the lever attachment screw on the plunger top is placed correctly in the track of the

gradient delivery system and then tighten the syringe holder screw. Attach the Tygon™ tubing to the three-way port and tighten the screw.

14. Repeat steps 11 and 12 with the low denaturing solution and syringe but place the syringe in the front of the gradient delivery system. When pouring a parallel denaturing gradient gel, the gel is top loaded and therefore the high solution is located at the back.
15. Start casting the gel by carefully rotating the wheel to deliver the solutions in a slow and steady manner. Make sure that the gel is not leaking out the bottom of the sandwiched glass plates and that the back screws are still sitting correctly in the tracks on the gradient delivery system.
16. Continue casting until the gel is about 5 cm away from the top of the front glass plate. Remove the taped needle and insert the appropriate well comb. Ensure that there are no bubbles in the wells and that the gel reaches the top of the front glass plate. The gel is now cast. Allow the gel to polymerize for 3 h.

5.4.1.1 Cleaning Procedures After Casting the DGGE Gel

17. After casting the gel, remove the syringes from the gradient delivery system and place the needle in a beaker of hot water. Use the syringes to pull water through all the tubing and prevent the solutions from polymerizing inside the tubing.
18. Rinse the syringes three times with hot water. Disconnect the tubes and use the syringes to flush them three times with hot water. Push air through each tube a number of times and invert to dry.
19. Rinse all other equipment (Falcon™ tubes, graduated cylinders, etc.) with hot water and invert to dry.

5.4.2 Preparing and Pre-Heating the Tank Buffer

1. About 1 h before loading the gel, the electrophoresis tank and heater needs to be turned on to warm the buffer to 60 °C.
2. Put together another set of sandwiched glass plates, but this time do not use spacers. Ensure that the plates are flush at the bottom and the sandwich clamps are tight. These plates are used at the back of the electrophoresis device (the core) to create the back half of the upper buffer chamber (**Note 18**).
3. To assemble the core, lay the device down with the back facing up. There is only one way the core fits into the electrophoresis tank; the red button on top of the core faces right while the black button faces left. To add the back glass plates, lay the core down so the black button is on the right.

4. Use 1× TAE and a pipette to wet the white U-shaped rubber seal. Turn the plates over so that the back plate is facing up. Align the pins on the side of the core with the grooves in the sandwich clamps. Insert the sandwiched plates at about 30° angle. Once inserted, press down near the bottom of the sandwich clamps until a click is heard.
5. Repeat with the sandwiched glass plates that contain the cast gel. Flip the core over (the red button should now be on the right hand side). Wet the white U-shaped seal with 1× TAE. Insert the glass plates with the shorter front plate facing down. Press down until a click is heard. Ensure that the top white seal formed, thereby creating the upper buffer chamber. Carefully remove the gel comb. Some of the wells may collapse but this problem is easily rectified later (step 10).
6. Add ~7 L of 1× TAE buffer to the electrophoresis tank. The buffer must be replaced after ~45 h, but for best results replace the buffer with each run. Fill the electrophoresis tank to at least the "fill" line indicated on the tank.
7. Carefully insert the core containing both sets of sandwiched glass plates. The red button should face the right-hand side.
8. Add ~350 mL of 1× TAE to the upper buffer chamber located between the two sets of sandwiched glass plates. Ensure that the upper buffer chamber is not leaking. If the seal is inadequate, the upper buffer chamber will not fill or maintain the buffer level. If this happens, simply remove the core and try reinserting both sets of sandwiched glass plates (a.k.a repeat steps 3–7).
9. Remove the heater from the lid stand and place it on top of the core. Ensure the stir bar sits in the hole at the bottom on the tank. Connect the power cord and turn the pump and heater on to test the buffer levels. Wait ~1 min and confirm the upper buffer chamber is full and holding a seal, and that the tank buffer level is at "max", as indicated on the electrophoresis tank. If necessary, add more buffer by removing the top clear loading panel. It is critical that both the heater and pump are completely submersed otherwise improper heating will occur.
10. To wash out the gel lanes and turn off and unplug the power supply. Remove the clear top loading panel, and insert a flat tipped pipette into the wells of the gel from this opening. The pipette tip must fit between the front and back glass plates. This takes a little practice but works best if you can sit with the gel at eye level. Wash the wells with 1× TAE to remove any un-polymerized denaturing solution. Also use the pipette tip to straighten out collapsed wells at this point. Once complete, add the clear top loading panel, connect the power

and turn on the pump and heater. Allow the system to heat to 60° (~1–1.5 h).

5.4.3 Loading and Running the DGGE gel

5.4.3.1 Loading the DGGE

1. Mix 35–40 μL of CPS-PCR products with ~7 μL of 6× loading buffer.
2. Turn off the heater and pumps and disconnect the power supply.
3. Remove the clear top loading panel on the heater unit.
4. Wash out wells in the gel again with 1× TAE. Use this time to straighten out and survey the wells to determine the best ones to load with PCR products. Do not use outside wells or any wells that are malformed.
5. Set pipette to ~50 μL over the volume to be loaded (85–90 μL, in this case). This will allow the wells to be washed once more before the samples are loaded.
6. Use a flat-tipped pipette tip to dispense samples as deeply into the wells of the gel as possible. The air in the pipette will wash out the well. Avoid any additional air bubbles once the sample starts to fill the well. Do this by walking the pipette tip up the well as the sample fills the well. Watch for the end of the sample and remove the pipette tip from the well before the final air bubble is released. Record the location of the sample and any additional comments.
7. Marker samples, used to normalize the gels for direct comparison, should flank the gel (**Note 19**). However, *do not* use the most outside lanes.
8. Once all the samples are all loaded, replace the clear top loading panel, connect the power supply, and turn the heater and pump on.

5.4.3.2 Running the DGGE

1. Attach the electrical leads to an appropriate DC power supply. Turn the electrodes on and set at 80 V.
2. Run the gel for 15 h at 80 V in 1× TAE buffer warmed to 60 °C.

5.4.4 Staining the Gel

5.4.4.1 Staining the DGGE

1. Add SYBR Green to the staining buffer. The staining buffer needs to be spiked each time it is used. Replace the buffer after 10–15 uses.
2. Upon completion of the run, turn off the electrodes, heater and pump. Disconnect the power supply. Wait 15 s. Remove the heater unit and place on the lid stand. Carefully lift the core and move it to a sink. Drain the top buffer chamber and remove the front set of sandwiched glass plates. Remove by pushing down on the black levers located on the sides of the core unit and gently pulling up.
3. Flip the glass plates so the shorter glass plate is facing up, and place on a clean absorbent surface. Carefully remove the sand-

wich clamps. Gently lift up on the spacers to help break the seal and remove the front glass plate. Remove the spacers taking care not to rip the gel. Cut a small amount of the gel off one corner; this serves as a marker for orienting the gel.

4. Submerge the back glass plate holding the gel into the SYBR Green staining solution. Place in a dark area such as a cupboard for > 3 h.
5. While the gel is staining, clean the glass plates and core device. Rinse the core device and sandwich clamps with hot water and invert to dry. Rinse the spacers with hot water and remove any gel pieces. Gently clean the glass plates with Liquid-nox and a scrub brush. Rinse with hot water and dry by spraying with 70% isopropanol.

5.4.4.2 Destaining the DGGE

1. After > 3 h of staining, destain the gel by moving it into 500 mL of Milli-Q. Destain for 30–60 min.

5.4.4.3 Viewing the DGGE

1. Visualize the gel using a transilluminator fit with a SYBR Green filter. Place a layer of water on the transilluminator and use the glass plate to move the gel from the destaining solution. Carefully slide the gel off the glass plate (make sure to clean the glass plate when finished). The gel is very thin and can tear easily. Use gloved hands and lots of water to move the gel around on the transilluminator and use the cut corner to orient the gel correctly. Photograph using a gel doc system.

5.5 Now what?

1. The gel can be discarded or bands cut out for sequencing.
2. Banding patterns and band intensity can be compared by importing the digital images of the gel into programs such as Gel Compar II (Applied Maths, St-Martens-Latem, Belgium).
3. Cut out bands of interest and place them in a sterile microfuge tube. Add 50–100 μL of 1× TAE and heat to 95 °C for 15 min to elute DNA. Store eluted DNA at −20 °C until ready to use.

6 Notes

1. The required volume of sample depends on the abundance of viral particles. Samples of a few liters are typically adequate for highly productive estuaries, while volumes of 100 L or more may be necessary for oligotrophic waters *(35)*. Approximately 10^9 viral particles (corresponding to ~50 ng of viral DNA) in a 10 mm well is necessary for a fingerprint *(21, 28)*. Assuming there are $\sim 10^7$ viral particles mL^{-1}, in theory the minimum sample volume is about 100 mL. However, viruses are typically lost during the concentration procedure; hence,

viral abundance should be determined at each step of the filtration and concentration procedures.

2. The glass-fiber filter will need to be changed frequently in productive waters.
3. Quantitative analysis of PFGE gels is aided by having virus standards of known genome sizes and abundances that are treated in the same way as the samples.
4. This step should be performed quickly or the agarose will solidify.
5. Always switch on the circulation pump before the cooling unit to prevent the buffer from freezing and damaging the cooling unit.
6. Carefully remove the plugs from the tubes, place them on a smooth surface, and dry them by touching the liquid film with the edge of a Kimwipe. Do not place the plugs on a Kimwipe or paper towel.
7. In our experience, this positioning of mass-ladders has proven to be useful.
8. The gel must be tightly connected to the platform otherwise the gel may detach from the platform and float freely in the electrophoresis chamber.
9. Electrophoresis conditions may vary. The values suggested here are for separating fragments from 10 to 200 kb. Increase pulse and run times to separate higher molecular weight size fragments (e.g., 1–15 s pulse ramp, 22 h *(5)*). Lower temperatures improve band resolution but increase the total run time *(5, 21, 36)*.
10. When closing the lid of the electrophoresis chamber, make sure that the electric connectors are tight, since they are easily unscrewed. Electrophoresis applies high voltage and/or power; always follow the safety recommendations of the equipment manufacturer.
11. This approach does not take into account the intensity of the bands; therefore, the abundance of viruses in each genome size is ignored.
12. Milli-Q can be used; however, to decrease contamination use ultrapure distilled, DNAse and RNase free, 0.1 μm filtered water (Gibco Invitrogen Corporation, Carlsbad, CA) when mixing up stock dNTPs and primers and when diluting these stock solutions. Also use this water in the PCR master mix.
13. Denaturing chemical recipe.
    Table 15.1 describes how to make the high and low denaturing solutions. It may be necessary to manipulate the percentage of denaturant or gel. Consult **Tables 15.2–15.4** below for the recipe needed to make other denaturing solutions.

14. The amount of DNA template used in a PCR is determined by the amount of DNA present in the initial sample. Ideally, ~100 ng of DNA should be added to a reaction. Depending on the question being addressed, the amount of DNA template added to any given PCR can be standardized by a spectrometric reading or by the initial sample volume.
15. If possible, avoid running a second round PCR. To increase sensitivity for DGGE, combine several independent first

Table 15.2
0% Denaturing solution but with different gel percentages *(37)*

	6% Gel	8% Gel	10% Gel	12% Gel
40% Acrylamide/bis (37.5:1, 2.6% C)	15 mL	20 mL	25 mL	30 mL
50× TAE	2 mL	2 mL	2 mL	2 mL
dH_2O	83 mL	78 mL	73 mL	68 mL
Total volume	100 mL	100 mL	100 mL	100 mL

Table 15.3
100% Denaturing solution but with different gel percentages *(37)*

	6% Gel	8% Gel	10% Gel	12% Gel
40% Acrylamide/bis (37.5:1, 2.6% C)	15 mL	20 mL	25 mL	30 mL
50× TAE	2 mL	2 mL	2 mL	2 mL
Formamide	40 mL	40 mL	40 mL	40 mL
Urea	42 g	42 g	42 g	42 g
dH_2O	To 100 mL	To 100 mL	To 100 mL	To 100 mL

Table 15.4
Use the following table to make denaturing solutions less then 100%. Use Acrylamide/bis, TAE, and water amounts outlined above in the 100% denaturing solution table (Table 15.3), but use the appropriate formamide and urea amounts listed below to achieve the desired amount of denaturant in the solution *(37)*

Denaturing solution	10%	20%	30%	40%	50%	60%	70%	80%	90%
Formamide (mL)	4	8	12	16	20	24	28	32	36
Urea (g)	4.2	8.4	12.6	16.8	21 g	25.2	29.4	33.6	37.8

Table 15.5
CPS primers and sequences

Primer	Sequence	Reference
CPS-1	5′-GTAGWATTTTCTACATTGAYGTTGG-3′	*(38)*
CPS-2	5′-GGTARCCAGAAATCYTCMAGCAT-3′	*(38)*
CPS-3	5′-TGGTAYGTYGATGG(A/C)AGA-3′	*(27)*
CPS-4	5′-CATWTCWTCCCAHTCTTC-3′	*(27)*
CPS-8	5′-AAATAYTTDCCAACAWATGGA-3′	*(27)*
CPS-9	5′-SWRAAATAYTTICCRACRWAGGATC-3′	*(26)*

Table 15.6
Size of PCR amplicons given the following primer pairs

Primers pairs	Size of PCR products	Reference
CPS-1 and CPS-2	165 bp	*(38)*
CPS-1 and CPS-4	430 bp	*(27)*
CPS-1 and CPS-8	592 bp	*(27)*
CPS-3 and CPS-4	860 bp	*(27)*
CPS-4 and CPS-9	592 bp	*(26)*

round PCRs. This helps increase the sample available for loading and may reduce PCR biases.

16. To ensure complete mixing of the high and low denaturing solutions, add a longer piece of tubing between the three-way port and the needle. Wrap the long tube around a cylinder for convenience and to enhance mixing (**Fig. 15.2**).
17. If using a different sized gel or spacer width adjust the volume setting according to the table (**Table 15.7**). The volume setting is located on the gradient delivery system above the syringe holder.
18. We recommend only running gels in the front of the electrophoresis tank, as the back glass plates are close to the heater which negatively affects gel quality.
19. Gels need to be normalized in order to compare band intensity and location among lanes and gels. Typically, this is done by flanking the gel with markers that have a distinct and repeatable banding pattern covering the spread of the gradient. An environmental sample can be used as long as there is enough to be loaded on every gel.

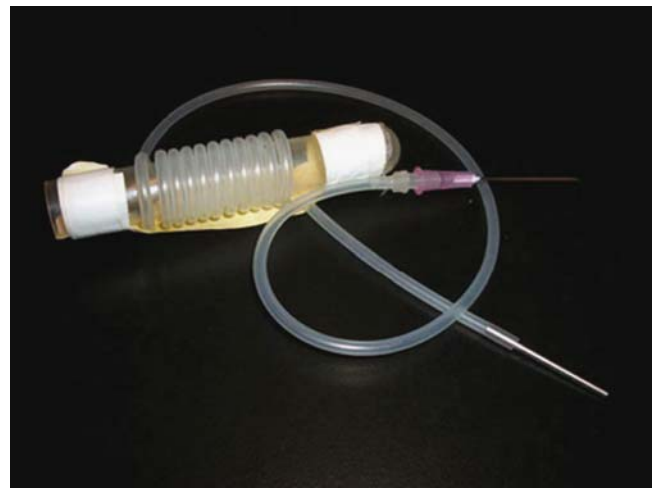


Fig. 15.2. Tygon™ tubing.

Table 15.7
Volume adjustment settings *(37)*

Spacers size	Gel size	Volume per syringe	Volume adjustment setting
0.75 mm	7.5 × 10 cm	5 mL	3.5
	16 × 10 cm	8 mL	6.5
	16 × 16 cm	11 mL	9.5
1.00 mm	7.5 × 10 cm	6 mL	4.5
	16 × 10 cm	11 mL	9.5
	16 × 16 cm	16 mL	14.5
1.5 mm	7.5 × 10 cm	8 mL	6.5
	16 × 10 cm	15 mL	13.5
	16 × 16 cm	24 mL	22.5

References

1. Suttle, C.A., *Viruses in the sea*. 2005. Nature, **437**(7057): 356–361.
2. Bergh, Ø., et al., *High abundance of viruses found in aquatic environments*. Nature, 1989. **340**: 467–468.
3. Proctor, L. and J. Fuhrman, *Viral mortality of marine bacteria and cyanobacteria*. Nature, 1990. **343**: 60–62.
4. Williamson, K.E., M. Radosevich, and K.E. Wommack, *Abundance and diversity of viruses in six delaware soils*. Applied Environmental Microbiology, 2005. **71**(6): 3119–3125.
5. Wommack, K.E., et al., *Population dynamics of chesapeake bay virioplankton: total-community analysis by pulsed-field gel electrophoresis*. Applied Environmental Microbiology, 1999. **65**(1): 231–240.
6. Cochlan, W.P., et al., *Spatial-distribution of viruses, bacteria and chlorophyll-a in neritic,*

oceanic and estuarine environments. Marine Ecology-Progress Series, 1993. **92**(1–2): 77–87.
7. Suttle, C.A., *Cyanophages and their role in the ecology of cyanobacteria*, in *The Ecology of Cyanobacteria*, B. Whitton and M. Potts, Editors. 2000, Kluwer Academic Publishers. 563–589.
8. Suttle, C., *The significance of viruses to mortality in aquatic microbial communities*. Microbial Ecology, 1994. **28**: 237–243.
9. Wilhelm, S.W. and C.A. Suttle, *Viruses and Nutrient Cycles in the Sea*. Bioscience, 1999. **49**(10): 781–788.
10. Middelboe, M., N. Jorgensen, and N. Kroer, *Effects of viruses on nutrient turnover and growth efficiency of noninfected marine bacterioplankton*. Applied Environmental Microbiology, 1996. **62**(6): 1991–1997.
11. Gobler, C.J., et al., *Release and bioavailability of C, N, P, Se, and Fe following viral lysis of a marine chrysophyte*. Limnology and Oceanography., 1997. **42**: 1492–1504.
12. Middelboe, M. and p.G. Lyck, *Regeneration of dissolved organic matter by viral lysis in marine microbial communities*. Aquatic Microbial Ecology, 2002. **27**(2): 187–194.
13. Fuhrman, J., *Marine vitruses and their biogeochemical and ecological effects*. Nature, 1999. **399**: 541–548.
14. Moebus, K., *Further investigations on the concentration of marine bacteriophages in the water around helgoland, with reference to the phage-host systems encountered*. Helgolander Meeresuntersuchungen, 1992. **46**(3): 275–292.
15. Sullivan, M.B., J.B. Waterbury, and S.W. Chisholm, *Cyanophages infecting the oceanic cyanobacterium Prochlorococcus*. Nature, 2003. **424**(6952): 1047–1051.
16. Nagasaki, K., et al., *Virus-like particles in Heterosigma akashiwo (Raphidophyceae): a possible red tide disintegration mechanism*. Marine Biology, 1994. **119**: 307–312.
17. Tarutani, K., K. Nagasaki, and M. Yamaguchi, *Viral impacts on total abundance and clonal composition of the harmful bloom-forming phytoplankton Heterosigma akashiwo*. Applied and Environmental Microbiology, 2000. **66**(11): 4916–4920.
18. Sahlsten, E., *Seasonal abundance in Skagerrak-Kattegat coastal waters and host specificity of viruses infecting the marine photosynthetic flagellate Micromonas pusilla*. Aquatic Microbial Ecology, 1998. **16**(2): 103–108.
19. Breitbart, M., et al., *Genomic analysis of uncultured marine viral communities*. Proceeding of the Natural Acadamy of Sciences, 2002. **99**(22): 14250–14255.
20. Steward, G.F., J.L. Montiel, and F. Azam, *Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments*. Limnology and Oceanography, 2000. **45**(8): 1697–1706.
21. Steward, G.F., *Fingerprinting viral assemblages by pulsed field gel electrophoresis (PFGE)*. In Methods in Microbiology, ed. J.H. Paul, Vol. 30, p. 666, 2001, San Diego: Academic Press.
22. Myers, R., T. Maniatis, and L. Lerman, *Detecting and localization of single base changes by denaturing gradient gel electrophoresis*. Methods Enzymol, 1987(155): 501–527.
23. Frederickson, C.M., S.M. Short, and C.A. Suttle, *The physical environment affects cyanophage communities in british columbia inlets*. Microbial Ecology, 2003. **46**(3): 348–357.
24. Short, S.M. and C.A. Suttle, *Use of the polymerase chain reaction and denaturing gel electrophoresis to study diversity in natural virus communities*. Hydrobiologia, 1999. **00**: 1–15.
25. Wommack, K.E. and R.R. Colwell, *Virioplankton: Viruses in Aquatic Ecosystems*. Microbiology and Molecular Biology Reviews, 2000. **64**(1): 69–114.
26. Short, C.M. and C.A. Suttle, *Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments*. Applied Environmental Microbiology, 2005. **71**(1): 480–486.
27. Zhong, Y., et al., *Phylogenetic diversity of marine cyanophage isolates and natural virus communities as revealed by sequences of viral capsid assembly protein gene g20*. Applied Environmental Microbiology, 2002. **68**(4): 1576–1584.
28. Suttle, C., A. Chan, and M. Cottrell, *Use of ultrafiltration to isolate viruses from seawater which are pathogens to marine phytoplankton*. Applied Environmental Microbiology, 1991. **57**: 721–726.
29. Short, S.M. and C.A. Suttle, *Sequence analysis of marine virus communities reveals that groups of related algal viruses are widely distributed in nature*. Applied Environmental Microbiology, 2002. **68**(3): 1290–1296.
30. Hennes, K., C. Suttle, and A. Chan, *Fluorescently labeled virus probes show that natural virus populations can control the structure of marine microbial communities*. Applied Environmental Microbiology, 1995. **61**(10): 3623–3627.
31. Noble, R.T. and J.A. Fuhrman, *Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria*. Aquatic Microbial Ecology, 1998. **14**(2): 113–118.

32. Wen, K., A.C. Ortmann, and C.A. Suttle, *Accurate estimation of viral abundance by epifluorescence microscopy*. Applied Environmental Microbiology, 2004. **70**(7): 3862–3867.
33. Williamson, K.E., K.E. Wommack, and M. Radosevich, *Sampling natural viral communities from soil for culture-independent analyses.* applied environmental microbiology, 2003. **69**(11): 6628–6633.
34. Chen, F. and C. Suttle, *Amplification of DNA polymerase gene fragments from viruses infecting microalgae*. Applied Environmental Microbiology, 1995. **61**(4): 1274–1278.
35. Chen, F., C. Suttle, and S. Short, *Genetic diversity in marine algal virus communities as revealed by sequence analysis of DNA polymerase genes*. Applied Environmental Microbiology, 1996. **62**(8): 2869–2874.
36. Birren, B. and E. Lai, *Pulsed field electrophoresis: A practical guide*. 1993, San Diego: Academic Press.
37. BioRad, *DCodeTm Universal Mutation Detection System instruction manual*. 1996, Hercules, Ca: BioRad.
38. Fuller, N.J., et al., *Occurrence of a sequence in marine cyanophages similar to that of t4 g20 and its application to PCR-based detection and quantification techniques.* Applied Environmental Microbiology, 1998. **64**(6): 2051–2060.

Chapter 16

Isolation Independent Methods of Characterizing Phage Communities 2: Characterizing a Metagenome

K. Eric Wommack, Shellie R. Bench, Jaysheel Bhavsar, David Mead, and Tom Hanson

Abstract

Current appreciation of the vast expanse of prokaryotic diversity has largely come through molecular phylogenetic exploration of sequence diversity within the universally conserved gene for small subunit ribosomal RNA (16S rDNA). A plethora of methodologies for characterizing the diversity and composition of bacterial communities is based on sequence polymorphisms within this single gene. By comparison, no gene is universally shared among viruses or bacteriophages, which has prevented broad scale characterization of viral diversity within microbial ecosystems. With the reduction in DNA sequencing costs and wide availability of bioinformatics software, the tools of whole genome shotgun sequencing are now beginning to be applied to the characterization of genetic diversity within whole microbial communities. Such metagenomic approaches are ideally suited to the characterization of natural assemblages of viruses, because of the typically small, coding-dense nature of viral genomes. Data from a limited number of characterized viral metagenome libraries within a range of microbial ecosystems indicates that viral assemblages are comprised of between ~1,000 to a million different genotypes. Furthermore, viral assemblages typically contain a large proportion of completely novel genes and are likely to be the largest reservoir of unexplored genetic diversity on earth. Here, we present a conceptual framework for characterization of viral assemblages through metagenomic approaches.

Key words: Nanoclone™, bioinformatics, database, sequence homology.

1 Introduction

In the decade, since the release of the first whole genome sequence for a bacteria *(1)* genomic tools and approaches have been applied to an ever broadening swath of microbiological diversity. For most of this short history, the focus has been on cultivated strains of bacteria and phage where the context

Martha R. J. Clokie, Andrew M. Kropinski (eds.), *Bacteriophages: Methods and Protocols, Volume 2: Molecular and Applied Aspects, vol. 502*, © 2009 Humana Press, a part of Springer Science+Business Media
DOI 10.1007/978-1-60327-565-1_16 Springerprotocols.com

of examining a whole genome integrates well with laboratory investigations of microbial biology. However, recent high profile investigations have shown that high-throughput sequencing can also provide a genetic snapshot of an entire microbial community and reveal otherwise unattainable insights on the potential biology of prokaryotes which are ecologically and phylogenetically distant from their cultivated cousins.

Community metagenome DNA sequence data is adding extraordinary detail to our view of bacterial and viral genetic diversity. For example, functional genes related to phototrophy once thought to be limited to narrow functional classes of bacteria are now known to be widespread among marine prokaryotes *(2)*. A recent metagenomic survey of bacterioplankon across a 4,000 m depth profile found that 10% of sequences could be attributed to cyanophage within water samples collected within the deep chlorophyll maximum zone *(3)*. As this survey utilized large insert phosmid libraries, it is likely that these sequences represent phage which were actively replicating within cyanobacterial populations. Access to these snapshots of microbial genetic diversity is opening new avenues for quantitative assessment of microbial community composition and characterization of the environmental relevance of particular gene functional groups *(4)*. Importantly, burgeoning metagenome sequence datasets provide the raw material necessary for constraining the evolutionary history of microbial life within an ecosystem context.

Based on the few metagenome surveys of marine viral assemblages completed to date, a consensus is emerging that virioplankton assemblages are extraordinarily diverse and contain an unusually high proportion of unknown (i.e., BLAST homolog to another environmental sequence of unknown identity) and completely novel sequences (i.e., sequences without a significant BLAST homolog) *(5)*. For long read length libraries (~700 bp per sequence), typically around 30% of viral sequences will have a BLAST homolog within a database of known genes (e.g., GenBank nr.). Unknown and novel sequences from viral metagenome libraries each comprise ca. 30% of the sequences, respectively. Against this backdrop of unknown and novel sequence, typically 60% of known sequences, to contain a collection of gene homologs that are relatively distant from better known representatives within bacterial genomes *(6, 7)*. For the remaining 40% of known sequences, typically the highest quality BLAST hits (i.e., lowest BLAST expectation (E) score) are to viral and prophage sequences. Despite the well-known predisposition for horizontal gene transfer between bacteriophage genomes *(8)*, there appears to exist a specific "marine" nature to viruses within the virioplankton as genes of known marine phages, cyanophages in particular *(9, 10, 12)*, are well represented within

virioplankton metagenome libraries. This finding is encouraging as it indicates that intensive study of a subset of marine bacteriophage may significantly advance understanding of autochthonous viral assemblages in the sea.

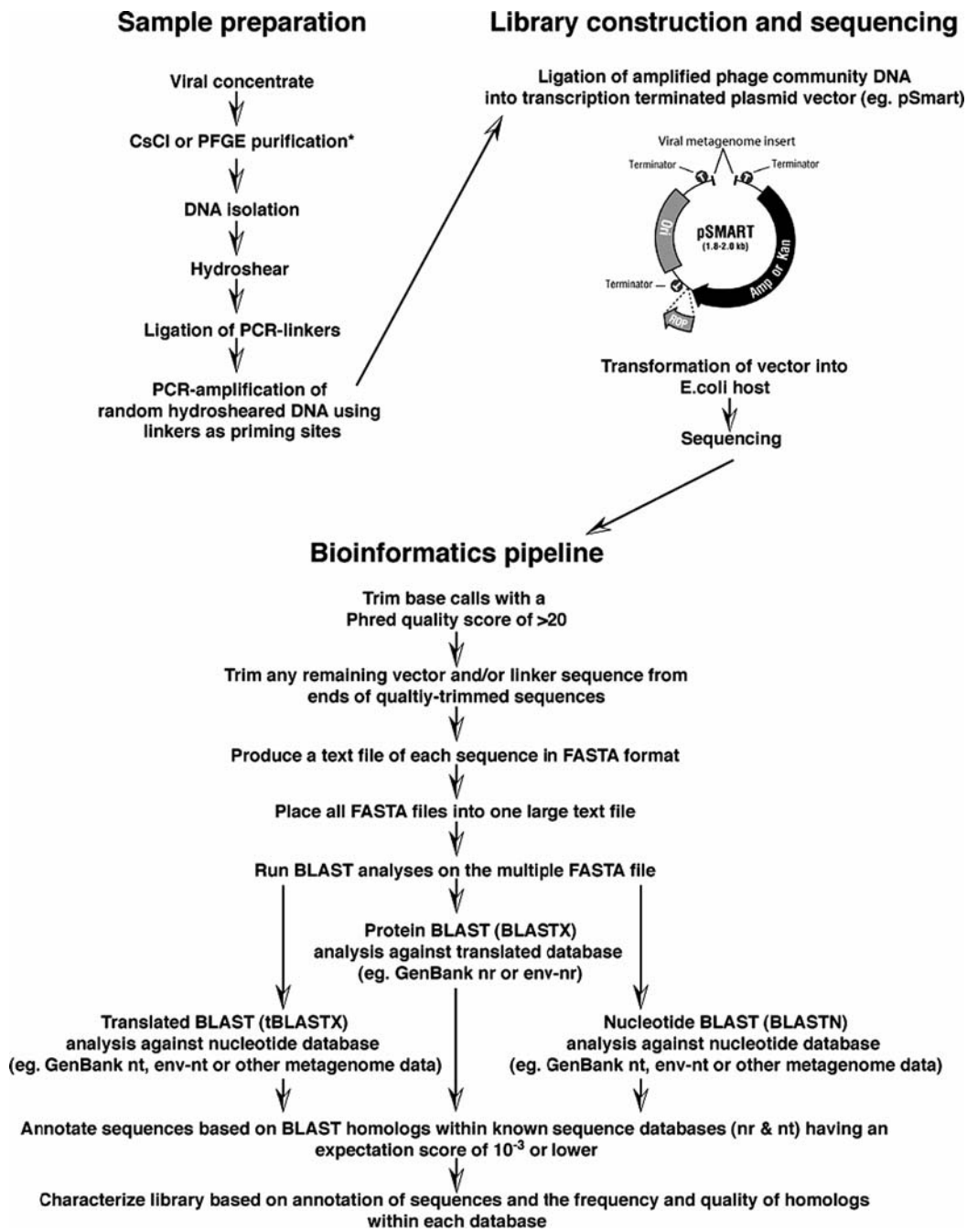


Fig. 16.1. Schematic diagram of steps involved in the description of a viral assemblage using metagenome sequence data obtained through traditional shotgun cloning and Sanger sequencing approaches.

Recent introduction of new sequencing technologies, which provide at least a 10-fold reduction in the cost of per base pair of DNA, promise to vastly improve access to high throughput DNA sequence data *(11)*. The one caveat of these technologies is their significantly shorter read length, ca. 100 bp, as compared to traditional Sanger sequencing (~700–900 bp). As demonstrated in a recently released metagenome dataset (http://scums.sdsu.edu/phage/Oceans *(9)*), short read sequences are of limited utility in describing the functional capabilities within the virioplankton. Nevertheless, this expansive dataset has proven extremely useful in testing hypotheses concerning the distribution of phage genotypes and the ecological principles governing the structure of virioplankton assemblages. Analyses based on assembly of short read sequences as well as homology searches against known phage genomes has clearly shown that virioplankton assemblages are extraordinarily even in composition, contain between 500 and 130,000 genotypes, and that significant overlap in genotype composition occurs between disparate geographic regions *(9)*. Because short read sequencing technology does not rely on a cloning step, these libraries revealed the presence of abundant gene homologs to a ssDNA phage (a chp1-like microphage) within a viral metagenome library from the Sargasso Sea.

Metagenomic characterization of a natural viral assemblage requires a number of technical approaches ranging from field collection and processing of samples, to shotgun library construction and bioinformatic analysis of DNA sequences, detailed description of each step is not feasible within the constrains of this chapter. Rather, this report will serve to briefly describe each step (**Fig. 16.1**) in the process of characterizing a viral assemblage through metagenome sequence data and direct the reader to resources for more in depth information.

2 Methods

2.1 Sample Preparation

An essential first step to metagenome characterization of a viral assemblage is obtaining a clean, high concentration sample of viral particles from an environmental sample. Approaches to purification and concentration of virus particles from environmental samples are detailed in **Chapter 1**. The primary challenges to construction of a representative viral metagenome library are obtaining adequate amounts of DNA for molecular cloning and ensuring that all viral genes, including those which are potentially lethal to the *E. coli* plasmid host, are represented in the library. Traditional methodologies for the construction of shotgun sequencing libraries require 10–100s of micrograms of purified DNA

depending on insert size. Obtaining such large amounts of DNA from virus particles within an environmental sample is difficult as it requires excessively large sample sizes. Assuming that the average genome size of a marine bacteriophage is similar to that of phage λ (48.5 kbp); and that typical viral abundance for a coastal water sample is $\sim 10^7$ viruses ml^{-1}, obtaining a 25 μg sample of virioplankton dsDNA would require concentration of virus particles from a 200 L water sample. This calculation assumes a 50% efficiency in both the concentration of virus particles and the subsequent purification of viral dsDNA, which is seldom the case. Inefficiencies and losses at every stage of the process typically result in recovery of 1–100 ng of pure viral DNA. While methodologies exist for collection and concentration of virus particles from large environmental samples, to date metagenome libraries constructed from viral dsDNA directly isolated from an environmental sample have not been reported, although recent technical progress demonstrates that it is feasible (DM, unpublished observations). In all cases, a small sample of viral dsDNA has been PCR-amplified prior to library construction. This amplification approach mitigates the necessity for large sample sizes and yields high purity dsDNA which is more amenable to cloning than an environmental sample. The anonymous DNA amplification approach of metagenomic library construction *(13)* is essentially the same as that utilized for clone free pyrosequencing *(11)*.

2.2 Library Construction

Because all commercially available plasmid sequencing vectors are propagated within an *E. coli* host, and clone libraries of viral metagenome sequence will contain a subset of genes which will be lethal to any bacterial host, it is essential that transcription of recombinant viral DNA be prevented. Typical blue/while sequencing vectors are designed to allow transcription and translation of insert DNA. As a consequence, libraries of viral metagenomic DNA cloned into typical sequencing vectors would be biased as lethal viral genes (e.g., genes associated with cell lysis) would not be represented. To circumvent this and other problems associated with difficult to clone DNA, commercially available vectors have been designed which prevent transcription/translation of insert DNA by eliminating the indicator *lac*Z gene and further protecting the plasmid by inclusion of transcription terminators adjacent to either end of the multiple cloning site (e.g., pSMART, Lucigen Corp., **Fig. 16.1**). Shotgun libraries of viral metagenomic DNA constructed through random shearing; ligation of short linker priming sites; PCR amplification from linker sites; and cloning in transcription-free vectors has been a highly successful approach to DNA sequence-based characterization of the composition and diversity of viral assemblages from

water samples and aquatic sediments *(7, 12)*, soil *(14)*, and mammalian feces *(15, 16)*. To circumvent the methodological hurtles of library construction from minute quantities (1–100 ng) of environmental viral DNA, many investigators have utilized contract cloning services such as Nanoclone provided by Lucigen Corp. (Middleton, WI) *(7, 16, 17)*. Once the challenge of library construction has been met, viral metagenomic clone libraries can subsequently be analyzed using high throughput DNA sequencing instrumentation along with freely available software for homology searches against public DNA sequence databases.

2.3 Bioinformatics Pipeline

The burgeoning availability of DNA sequence data has fueled vibrant development of computer software and tools (e.g., scripts and databases) for the analysis of sequence data. New tools, software, and analytical approaches are constantly appearing in this fast moving field; thus, the following methodological overview should be only regarded as an outline for the analysis viral metagenome sequence data.

Initial base calls for each clone sequence will include low quality and indeterminate bases (Ns) as well as contaminating vector and, in the case of nanoclone libraries, PCR linker sequence. Accurate analysis of viral metagenome sequence requires that poor quality base calls and contaminating, non-viral, sequence be removed prior to bioinformatic characterization of sequences. The most commonly used suite of applications for cleaning DNA sequences is the Phred, Phrap, Consed suite *(18, 19)* which is freely available to academic and non-profit users (http://www.phrap.org/) and runs on Mac OS X, a number of UNIX operating systems as well as Microsoft Windows. Using a quality score of ≥ 20, chromatogram files from Applied Biosystems, Molecular Dynamics or LiCor sequencers are processed by Phred producing a FASTA formatted output sequence file. Subsequently, contaminating sequence (vector and linker) can be screened from the FASTA files using Cross_Match which is part of the Phred, Phrap, Consed suite. Cleaned viral metagenome FASTA formatted sequences should be placed sequentially into a single text file for BLAST analyses. Because many of the steps in bioinformatic analysis of DNA sequences require batch editing or parsing of text files, a working knowledge of a scripting language such as PERL can be extraordinarily helpful.

Metagenomic characterization of a viral assemblage is primarily based on describing the sequence library according to the information gained through homology searches using versions of the Basic Local Alignment Search Tool (BLAST) *(20)*. While there are several BLAST program versions which run on the

complete range of modern operating systems, conceptually there are only four possible approaches to homology searches based on primary genetic sequence data: nucleotide query sequence versus nucleotide sequence database (BLASTN); protein query sequence versus protein sequence database (BLASTP); translated nucleotide query sequence versus protein sequence database (BLASTX); and translated nucleotide query sequence versus translated nucleotide sequence database (tBLASTX). Because of degeneracy in the amino acid code and the large pool of unknown sequences within most viral assemblages, nucleotide searches of viral metagenome sequences are less likely to find significant genetic homologs to known sequences. Instead, BLAST searches based on identity of amino acids rather than nucleotides [i.e., translated viral metagenome sequences against protein databases (BLASTX) or translations of nucleotide databases (tBLASTX)] are the preferred means of finding homologs which are based on true functional similarly between genes. All viral metagenome studies to date have selected a BLAST expectation (E) score of $< 10^{-3}$ or 10^{-5} as the cut-off point for determining the significance of a particular BLAST hit.

The largest collection of sequence databases available for BLAST homology searches are within the GenBank (http://www.ncbi.nlm.nih.gov/Genbank/index.html) databases: nucleotide (nt, $\sim 4.6 \times 10^6$ sequences), non-redundant protein (nr, $\sim 4.2 \times 10^6$ sequences); environmental nucleotide (env-nt, $\sim 1.5 \times 10^6$ sequences); and environmental protein (env-nr, $\sim 1 \times 10^6$ sequences). However, the all-inclusive nature of GenBank sequence databases, can create challenges for subsequent analyses based on sometimes poorly annotated BLAST homologs. Thus, smaller boutique databases focusing specifically on viral genomic and metagenomic sequence data (e.g., http://www.viralecology.org and http://scums.sdsu.edu) or carefully curated sequence databases such as SEED *(21)* can assist in downstream characterization and annotation of viral metagenome sequences.

For most viral metagenome studies, the approach to library characterization has been to cast a broad search for homologous sequences using BLASTX and tBLASTX homology searches against the four GenBank databases mentioned above as well as smaller viral metagenome databases. If each sequence is compared to at least four databases and the top five BLAST homologs for each database are recorded, then, it is possible that a single sequence will have as many as 20 homologous sequences. Because, even small metagenome libraries of a few hundred sequences can pose a significant analytical challenge, a critical component to effective description of a metagenome library is a database for storage and query of BLAST results (**Fig. 16.2**).

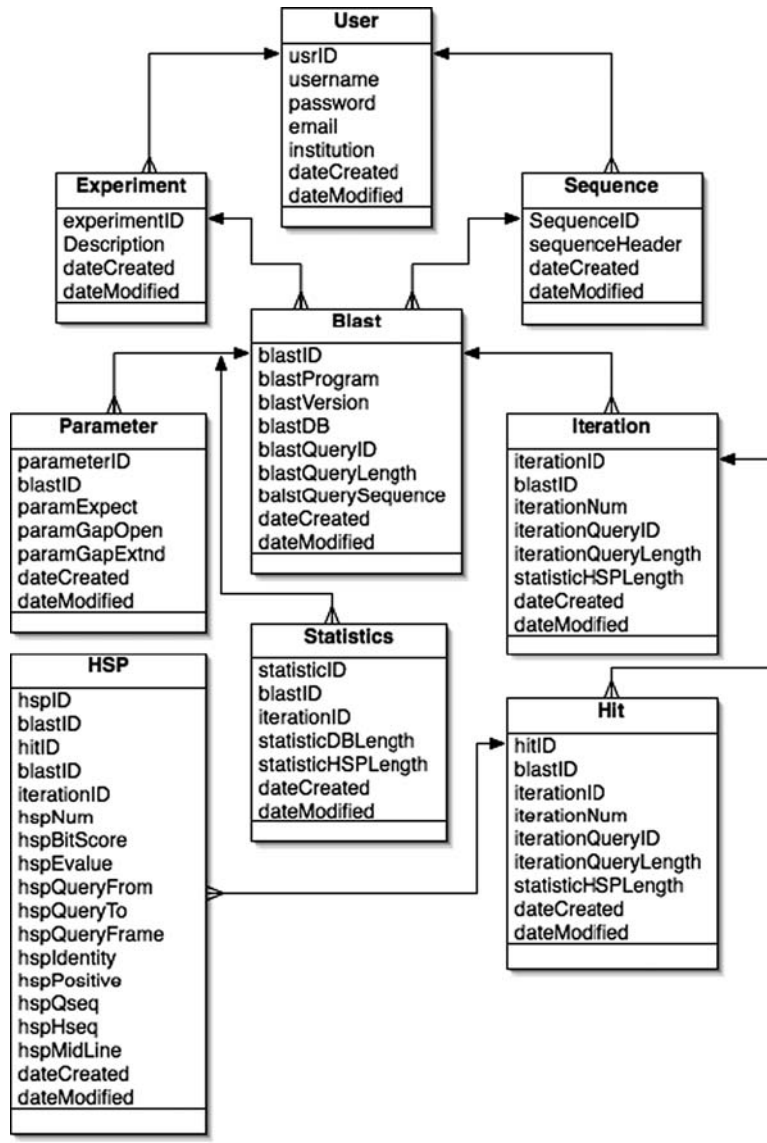


Fig. 16.2. Simplified database schema for the storage and analysis of BLAST search results for viral metagenome sequence data.

Several commercial (e.g., Microsoft Access, Filemaker, and Oracle) and open source (e.g., MySQL) database applications can easily handle the database requirements for a typical viral metagenome study. The choice of application depends in part on the size of the database, data processing requirements, and the need for flexibility in building a user interface for interaction with the database. Because of its flexibility, publicly available open source code, and ability to handle very large datasets MySQL (http://www.mysql.com/) is the most commonly used

database application within the bioinformatics community. Moreover, a wealth of open source graphic user interface tools are available (e.g., phpMyAdmin, http://www.phpmyadmin.net) for design, and administration of MySQL databases. Alternatively, commercial database applications, MS Access and FileMaker in particular, while offering less flexibility can be more approachable for the non-specialist.

In the example MySQL database schema, the table structure and many of the field codes have been directly imported from the DTD (document type definition) of BLAST XML (extensible mark-up language) output (**Fig. 16.2**). Once the database is designed and constructed, BLAST results, recorded either at text or at XML files, can be directly parsed into the tables through the use of scripts. Subsequent tasks of characterizing viral metagenome sequences are accomplished in large part through queries against the BLAST results database. Here again, some working knowledge of database structure, a scripting language, and SQL (structured query language) is critical to success.

Description of a viral metagenome library typically encompasses the frequency of significant BLAST hits (i.e., homologs) within taxonomic groups (e.g., eukaryotes, eubacteria, archaea, and viruses) and functional classes of proteins and stable RNAs. The hierarchy of descriptors used for functional classification varies from study to study; however, there are several schemes available including pFAM *(22)*, Gene Ontology *(23)*, and TIGRFAM *(24)*. In addition to the familiar Linnean-style classification of viruses maintained by the International Committee for the Taxonomy of Viruses (ICTV *(25)*), we have also used a bacteriophage classification scheme based on commonalities in gene content known as the phage proteomic tree *(26)*. Mobile genetic elements (prophage, transposons, and plasmids) are perhaps the most difficult category of BLAST homologs to characterize within a viral metagenome sequence library. Some of these hits can be quickly annotated based on the identity of the homolog; while others require more detailed inspection of the genetic context of the original BLAST homolog. Our approach has been to look for viral metagenome sequences which have BLAST homologs to hypothetical proteins within bacterial whole genome sequences. Once this subset of viral metagenome sequences is identified, neighboring genes of the original hypothetical protein homolog are inspected to determine if they are part of a mobile genetic element residing within the bacterial genome. In the case of prophage, oftentimes the hypothetical protein BLAST hit is found within a 25 kb or larger collection of syntenic genes which include capsid structural genes or gene associated with the establishment or maintenance of lysogeny (e.g., integrase and repressor).

Acknowledgements

The authors gratefully acknowledge the support of the National Science Foundation Microbial Observatories program (grant number MCB-0132070 awarded to K. E. Wommack) and the assistance of a USDA National Needs Graduate Fellowship to S.R. Bench.

References

1. Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269: 496–512.
2. Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74.
3. DeLong, E.F., Preston, C.M., Mincer, T., Rich, V., Hallam, S.J., Frigaard, N.U., Martinez, A., Sullivan, M.B., et al. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. Science 311:496–503.
4. Tringe, S.G., von Mering, C., Kobayashi, A., Salamov, A.A., Chen, K., Chang, H.W., Podar, M., Short, J.M., et al. (2005) Comparative metagenomics of microbial communities. Science 308:554–557.
5. Edwards, R.A. and Rohwer, F. (2005) Viral metagenomics. Nat Rev Microbiol 3: 504–510.
6. Breitbart, M., Felts, B., Kelley, S., Mahaffy, J.M., Nulton, J., Salamon, P., and Rohwer, F. (2004) Diversity and population structure of a near-shore marine-sediment viral community. Proc R Soc Lond B Biol Sci 271:565–574.
7. Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D., Azam, F., and Rohwer, F. (2002) Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci U S A 99:14250–14255.
8. Hendrix, R.W., Smith, M.C., Burns, R.N., Ford, M.E., and Hatfull, G.F. (1999) Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc Natl Acad Sci U S A 96:2192–2197.
9. Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., Carlson, C., Chan, A.M., Haynes, M., et al. (2006) The Marine Viromes of Four Oceanic Regions. PLoS Biol 4:e368.
10. Paul, J.P. (2006) Personal communication.
11. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380.
12. Bench, S.R., Hanson, T.E., Williamson, K.E., Ghosh, D., Radosovich, M., Wang, K., and Wommack, K.E. (2007) Metagenomic characterization of Chesapeake Bay virioplankton. Appl Environ Microbiol 73:7629–7641.
13. Schoenfeld, T., Patterson, M., Richardson, P.M., Wommack, K.E., Young, M., and Mead, D. (2008) Use of a novel cloning strategy for comparative metagenomic analysis of viral assemblages from Yellowstone Hot Springers. Appl Environ Microbiol 74:4164–4174.
14. Wommack, K.E., Bench, S.R., Williamson, K.E., and Radosevich, M. (2004) Viruses in soils: The first terrestrial viral metagenome, *In* 10th International Symposium for Microbial Ecology, Cancun, Mexico
15. Breitbart, M., Hewson, I., Felts, B., Mahaffy, J.M., Nulton, J., Salamon, P., and Rohwer, F. (2003) Metagenomic analyses of an uncultured viral community from human feces. J Bacteriol 185:6220–6223.
16. Cann, A.J., Fandrich, S.E., and Heaphy, S. (2005) Analysis of the virus population present in equine faeces indicates the presence of hundreds of uncharacterized virus genomes. Virus Genes 30:151–156.
17. Breitbart, M. and Rohwer, F. (2005) Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. BioTechniques 39:729–736.
18. Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186–194.
19. Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175–185.

20. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
21. Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.Y., Cohoon, M., de Crecy-Lagard, V., Diaz, N., et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 33:5691–5702.
22. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., et al. (2002) The Pfam protein families database. Nucleic Acids Res 30: 276–280.
23. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.
24. Haft, D.H., Selengut, J.D., and White, O. (2003) The TIGRFAMs database of protein families. Nucleic Acids Res 31:371–373.
25. Büchen-Osmond, C. (2003) The universal virus database ICTVdB. Computing in Science and Engineering 5:16–25.
26. Rohwer, F. and Edwards, R. (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. J Bacteriol 184:4529–4535.

Section IV

Applied Aspects of Bacteriophage Biology

Chapter 17

Phage Typing

Irina Chirakadze, Ann Perets and Rafiq Ahmed

Abstract

Phage typing is a rapid, economical, reliable, and reproducible technique, requiring no specialized equipment, for fingerprinting disease-causing agents for epidemiological investigation and surveillance.

Key words: Phage typing, lysotyping, biotyping, ribotyping, PFGE.

1 Introduction

An increase of epidemics, epizootics, and nosocomial (hospital-acquired) infections caused by pathogenic and conditionally pathogenic bacteria is being observed in developing and developed countries all over the world. These diseases often display an important polymorphism of clinical picture, such as asymptomatic, chronic, sub-clinical, atypical, different types of carriers—healthy and post-infection. For epidemiological purposes, simply that a patient is suffering from an infection by *Salmonella enterica* serovar Typhimurium is often insufficient. This information provides us with the genus and species but we often need a fingerprint of the isolate so that the source of the infection can be tracked which will permit the public health authorities to intervene by removing contamination sources from circulation, preventing the population from exposure to contamination, and controlling further spread of the disease. A wide

This chapter is dedicated to Dr. Elena Makhashvili, who worked closely with George Eliava and Felix d'Herelle in establishing what is now the Eliava Institute, took over leadership when Eliava was executed by Beria in 1937, and established the All Soviet Phage Typing Centre there in Tbilisi, in 1964.

Martha R. J. Clokie, Andrew M. Kropinski (eds.), *Bacteriophages: Methods and Protocols, Volume 2: Molecular and Applied Aspects, vol. 502*, © 2009 Humana Press, a part of Springer Science+Business Media
DOI 10.1007/978-1-60327-565-1_17 Springerprotocols.com

variety of molecular and physiological tools have been employed for characterization at the strain or isolate level *(1, 2)*.

Intraspecies differentiation of bacteria can be based on taxonomic features, such as morphology, biochemical properties (biotyping), virulence (pathotyping), and antigenic structure (serotyping). In addition, a wide variety of genome-based taxonomic techniques *(1)* have been developed such as pulsed-field gel electrophoresis (PFGE) *(3)*, amplified-fragment length polymorphism (AFLP) *(4)*, and amplification of repetitive bacterial DNA elements (REP-PCR) *(5)*. Other typing systems are based on sensitivity to specific chemicals, including antibiotics, plasmid profiling, ribotyping, and the production of or sensitivity to bacteriocins and bacteriophages (lysotyping or phage typing). Phage typing provides long-term and internationally comparable surveillance data. Because of their recent introduction, such information is not available for molecular techniques. Phage typing of enteric pathogens has been and still is being used successfully to characterize disease-causing agents for epidemiological investigation and surveillance *(6–8)*. Small laboratories may have difficulties in maintaining expertise in phage typing. However, participation in quality control programs like Enter-Net External Quality Assurance, which has been in place for the last 15 years, may be helpful in alleviating such problems. Phage typing is a rapid, economical, reliable, and reproducible technique requiring no specialized equipment *(9)*.

In 1922, B.R. Callow *(10)* was the first to demonstrate that a panel of staphylococcal phages differed in their lytic activity against host cells *(11)*. The first phage typing for a Gram-negative bacterium, *Salmonella* Typhi, was developed in 1938 as a result of the work of Craigie and Yen on Vi-specific phages *(12–14)*. Many of the early studies employed lysis of broth cultures as a measure of sensitivity. Fisk *(15)* introduced the plate lysis technique which is the basis of all modern phage typing systems. Phage typing of bacterial species is an important method for epidemiological diagnostics and has been applied to a great variety of bacterial genera (with emphasis on first publications) including: *Bacillus (9,16), Burkholderia (17), Campylobacter (18), Clostridium difficile (19, 20), Corynebacterium (19, 21), Enterobacter (22), Escherichia (23–26), Listeria (27)*, *Mycobacterium (28, 29), Pasteurella (30), Proteus (31, 32), Pseudomonas (33, 34), Salmonella (12, 13, 35–38), Serratia (39), Shigella (40), Staphylococcus (10, 11, 15), Streptococcus (41–43), Vibrio (44, 45), Yersinia (46)*, etc. Two of the advantages of this procedure are that the results are rapidly acquired, and the associated costs are relatively low.

Multiple phage typing schemes have been developed for the same bacterium. For example is the case of *Salmonella enterica* serovar. Typhimurium schemes have been elaborated by Adlakha

(47), Anderson *(37, 48, 49)*, Callow *(36)*, Felix *(50)*, Ibrahim *(51)*, Lilleengen *(52)* and Wilson *(53)*. At present there are over 200 definitive phage types (DTs) of *S. Typhimurium (54)*.

A functional phage typing system includes the following characteristics:

a) A panel of genetically and phenotypically stable temperate or lytic phages possessing broader rather than narrow host-range specificities.

b) Results, which are obtained quickly and are clear-cut and require limited training in interpretation.

c) Method that can be standardized.

d) Bacterial cells must display a stable phage type over time.

e) Bacterial cells should form a lawn upon which phage lytic reactions can be easily determined. Hosts such as certain *Pseudomonas aeruginosa* strains which display an autoplaque phenomenon are problematic in these regards and media should be devised to minimize autoplaquing, which presumably is associated with the induction of prophages that can plate even on the lysogenic strain from which they arise.

Definition: RTD = routine test dilution which is the highest dilution (i.e., lowest titer) which results in confluent lysis; termed the critical test dilution by Craigie and Yen *(12, 13)*.

European and North American phage typing of enteric bacterial pathogens is coordinated through the "International Federation of Enteric Phage Typing" at the Health Protection Agency Centre for Infections (61 Colindale Ave., London NW9 5EQ; http://www.hpa.org.uk/cfi/lep/sru.htm) and at the National Microbiology Laboratory, Public Health Agency of Canada (1015 Arlington Street Winnipeg, Manitoba R3E 3R2; http://www.phac-aspc.gc.ca & http://www.nml.ca), which collaboratively supply all WHO- and Enter-net-associated laboratories with phage concentrates originated at their centers.

Phage typing is performed by using a specific set of typing phages and corresponding phage typing schemes for specific species such as *Staphylococcus aureus* or serotypes as in *Salmonella*. Phage types included in a scheme differ from each other by one or more lytic reactions and provide different lytic patterns on different groups of bacterial strains. Each phage typing reaction is specific, repeatable, and stable for years for a strain kept intact at −80 °C. Each lytic pattern is considered a different phage type and is labeled with a unique identifier. The identifiers may be numbers or a combination of numbers and letters denoting slight association or complete uniqueness among phage types.

Isolation and selection of diagnostic typing phages for interspecies differentiation is termed phage typing. This method is used for the determination of the source of infection, the route of infection or disease transmission, outbreaks, epidemics,

epizootics, hospital or nosocomial infections, determination of sporadic cases or healthy post-infection carriers, study of distribution, and migration of phage types. It is also important in developing therapeutic phage cocktails.

Phage typing is very important for theoretical medicine. According to Scholtens, phage type is not only a set of bacterial strains with the same phage sensitivity, but strains with the same features of lysogeny when the typing set includes relevant temperate phages.

1.1 Type-Specific Phages for Diagnostics or Typing Phages

The typing phages may be isolated either from the environment (sewage, river, lake water, etc.) or from particular lysogenic bacterial strains. Phage isolation from the bacterial culture is possible directly from a rapidly growing broth culture, after UV irradiation or mitomycin-c treatment of the bacteria to induce prophages.

For the purposeful selection of typing phages, it is necessary to include phage-resistant and epidemiologically marked bacterial strains in the set of type strains. (Epidemiologically marked strains are epidemiologically related strains by time, place, and common source of infection.) These strains can be isolated from outbreak-associated humans, foods, animals, bacterial carriers, nosocomial infection source, outbreak, and sporadic cases, etc.

Biological features of the phages are a very important issue. Lytic as well as temperate bacteriophages may be used for typing.

1.2 Phage Propagation

The phages are propagated or cultivated in liquid media or on solid agar on their specific propagating strain. After propagation, each phage preparation is tested to verify its lytic reaction and specificity on a set of phage typing control strains. The phage titer is determined on corresponding propagating strains. The recommended titer for typing phage lysates is 10^7–10^9. (Undiluted concentrates may give false positive results due to the release of bacteriocins and/or other biological material present in phage suspensions by causing cell lysis.) The routine test dilution (RTD) or optimal titer provides the most clear and stable lytic reaction by diluting out potential interfering biological materials in the phage lytic reaction.

1.3 Necessary Features

Typing phages should demonstrate high lytic activity, high specificity (type ability), and high lytic reaction stability. The phages should exhibit a high level of discrimination, which means that each phage should reveal a specific lytic reaction. The lytic reaction may be evaluated using the following criteria: CL, confluent lysis, SCL, semi-confluent lysis, OL, opaque lysis, i.e., secondary growth of bacterial colonies. The reaction during phage typing should correspond to the reaction presented in the scheme.

Typing phages may be kept at 4 °C in syringes or in ampoules. Some phages lose their activity rather soon, but the majority of phages preserve activity over long periods of time.

1.4 Bacterial Strains

The study and control of specificity of the propagating host strains are important for the reference laboratories.

1. Propagating strains: typing phages (TP) are propagated on these strains. Each set of TP has its own particularly appropriate propagating host, in which it grows to high titer. If the strains are lysogenic, which is characteristic of many bacterial species such as *Salmonella, Pseudomonas, Staphylococcu*s (etc.), recombination between the typing phage and prophage occurring in the bacteria may happen. TP may change their specificity as a result of recombination. The propagating strain should be renewed from frozen stocks every 2–3 months. Propagating strains should not be changed because that may change the genetic profile of the typing phage.
2. Phage types: as was mentioned above; each phage type has a unique and specific lytic reaction on the related strain. All phage types included in the scheme must be presented with at least—two to three bacterial strains with different epidemiological patterns.
3. The specificity and taxonomic features of the phage types should be checked from time to time. A control set of phage-type bacterial strains are used for checking the specificity of the typing phage each time they are grown. The control set is used in case of introduction of a new typing phage in the scheme.

The following procedure is based on the standard operating procedures used by the Laboratory for Foodborne Zoonoses for *E. coli* O157:H7 and *Salmonella*. Using your own panel of phages, you also should be able to set-up a host-specific phage typing system.

2 Materials

2.1 Luria Bertani Agar, Miller (LB) (Difco Laboratories)

2.2 Sterile phage broth (*see* **Note 1**)

2.3 Phage agar:

Nutrient Broth (Difco)	160 g
NaCl	68 g
Bacto agar (Difco)	104 g
Distilled water	8000 mL
	pH 6.8

Melt, dispense in 500 mL screw-capped bottle, and autoclave.

2.4 We use a set of (*Salmonella*) typing phages diluted to RTD and kept in 3 mL syringes equipped with blunt canulas, contained in a dispenser plate holder. The syringes have a Luer-Lok Tip from Becton Dickinson Biosciences Ref #309585 (San Jose, CA; http://www.bdbiosciences.com/) and the blunt cannulas are 18 Gage ×1" (2.54 cm) Kendall Monoject Life Shield Blunt Cannula Ref # 8881202017 from Tyco Healthcare Products Company (Mansfield, MA; http://www.tycohealthcare.com/).

2.5 Phage broth culture of host bacteria.

2.6 Square Petri plates (100 × 100 × 15 mm)—Fisher Scientific.

2.7 **Equipment**:

1) Incubators set at 37 °C.
2) Aluminum syringe dispenser (Specialalloy, Winnipeg, Manitoba, Canada). This apparatus is based on the equipment described by Pruneda & Farmer *(55)*. (The dispensing can also be simply carried out manually.)

3 Methods

3.1 Preparation of Agar Plates

1. Pour about 40 ml of sterile molten phage agar into each plate (*see* **Note 3**).
2. When cool, repackage in the plastic sleeve and store at 4 °C. These agar plates may be used for up to 1 month.
3. Prior to use, gently dry agar plates either in an incubator set at 37 °C, or in a biological safety cabinet, with lids partially ajar, until a slight rippling is visible on the surface of the agar (ca. 1 h) (*see* **Note 4**).

3.2 Preparation of Bacterial Culture for Testing

1. Sub-culture each strain of the species (*Salmonella*) to be phage typed onto an LB agar plate to obtain isolated colonies.
2. Incubate the LB plate at 37 °C overnight.
3. Pick, with a sterile inoculating needle, a small amount from the center of a smooth colony and suspend each pick in one tube of LB broth.
4. Incubate these culture tubes in a water bath shaker at 37 °C and 100 rpm for approximately 1 h or until the culture becomes slightly turbid.
5. Flood a dried phage agar plate with the phage broth culture using a disposable transfer pipette to produce a bacterial "lawn" of the test strain.
6. Ensure that all the surplus broth culture is removed from the agar surface by tilting the agar plate and sucking the broth up at the lowest point with a sterile Pasteur pipette.
7. Allow the plate to dry thoroughly covered on the bench top for approximately 10 min.

3.3 Phage Inoculation onto the Test Strains Using a Multiplex Dispenser (the Phage Drops Can Also Be Dispensed by Hand Using a Pipetteman)

1. Take the dispenser plate holder containing syringes of the appropriate typing phages, at RTD, out of the refrigerator.
2. Assemble the system as shown in **Fig. 17.1**.
3. Place the dried plate containing the bacterial lawn in the apparatus (**Fig. 17.2**).
4. Dispense one drop of each phage on to the phage agar plate (*see* **Note 2**).
5. Remove phage agar plate and allow phage drops to dry on the phage agar plate for a maximum of 15 min with the lid ajar.
6. Repeat the process with a plate that has not been inoculated with bacteria. This is the sterility check on the phage preparations.
7. Return the dispenser plate holder containing the syringes to the refrigerator (**Fig. 17.3**).
8. Invert and incubate the dried plates at 37 °C overnight.
9. Examine the plates the next day; enter the reaction of each phage into a chart. The plates are read with the naked eye or a ×10 hand lens through the bottom of the plates using direct and oblique illumination (**Fig. 17.4**). The phage typing designation of the test culture is completed by careful comparison of the lytic reactions with the given reactions in the scheme table. The following is a very much-simplified description of the reactions:

Reaction – descriptive	Explanation
CL	Clear lysis in drop area—expected to generally be obtained with lytic phage
OL	Opaque or turbid lysis pattern—expected to be obtained with temperate phage
IPC	Individual clear plaques
IPO	Individual turbid (opaque) plaques
–	No lysis

10. The sterility plates should show no growth.

3.4 Anomalous Lytic Reactions

1. Atypical lytic reactions may occur due to the fact that the strain is undergoing variation due to the environmental stress on the strain during transportation, storage, freezing, or repeated sub-culturing, in return producing clones of varying degrees of type specificity. Testing 3–10 single colonies and finding ones with a typical lytic reaction can remedy this problem.

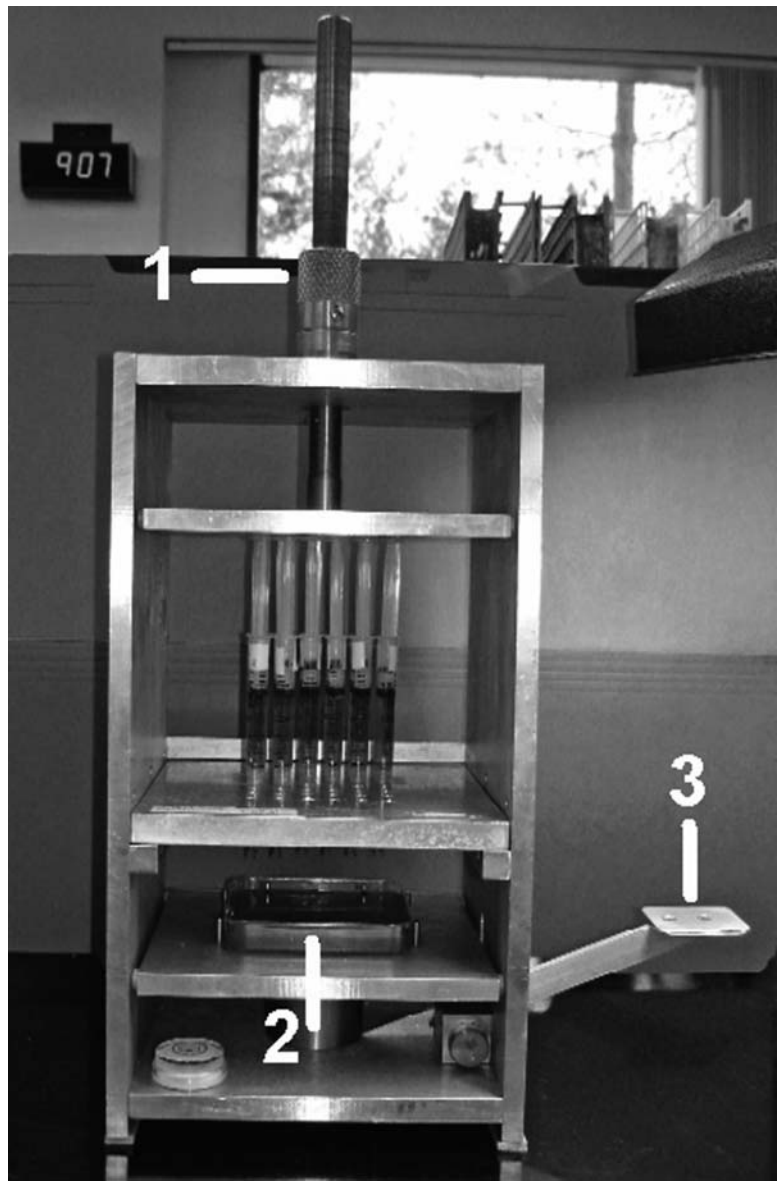


Fig. 17.1. Dispenser with phage-loaded syringes in place. Turning the knurled knob at the *top* (arrow 1) depresses the syringe plunger allowing small drops (*see* **Fig. 17.2**) to fall onto the plate (arrow 2). This is raised just below the cannulas by depressing the leaver arm (arrow 3).

2. New phage type–atypical reaction pattern of an outbreak or a large number of sporadic strains showing identical atypical lytic pattern may be due to a new phage type.

3.5 Phage Typing and Outbreak Investigation

Phage typing and PFGE are the two methods of choice readily employed by the laboratory workers for outbreak investigation and differentiation of outbreak strains from simultaneously

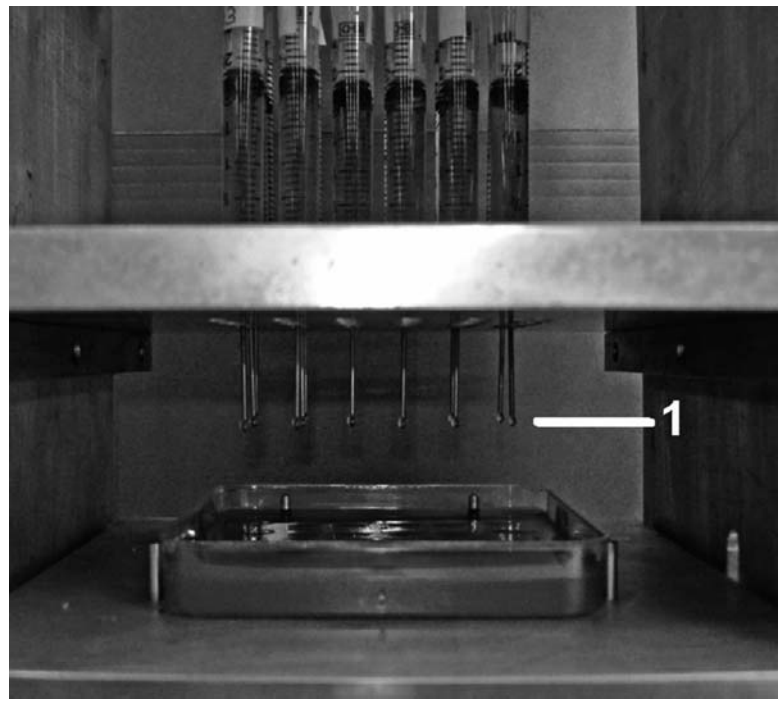


Fig. 17.2. Close-up of the drops of phage lysates (arrow 1) at the tips of the cannulas.

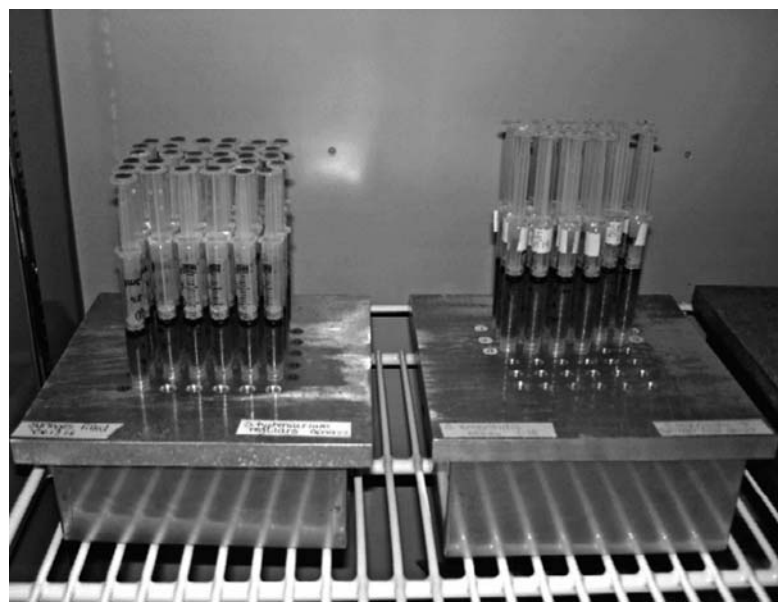


Fig. 17.3. Storage of phage-containing syringes in refrigerator.

occurring sporadic cases. Inherent shortcomings of Phage typing and PFGE arise from having a common or predominant phage type(s) and PFGE patterns in a particular geographical area—i.e., *S.* Heidelberg—and in some cases, a limited number of PFGE

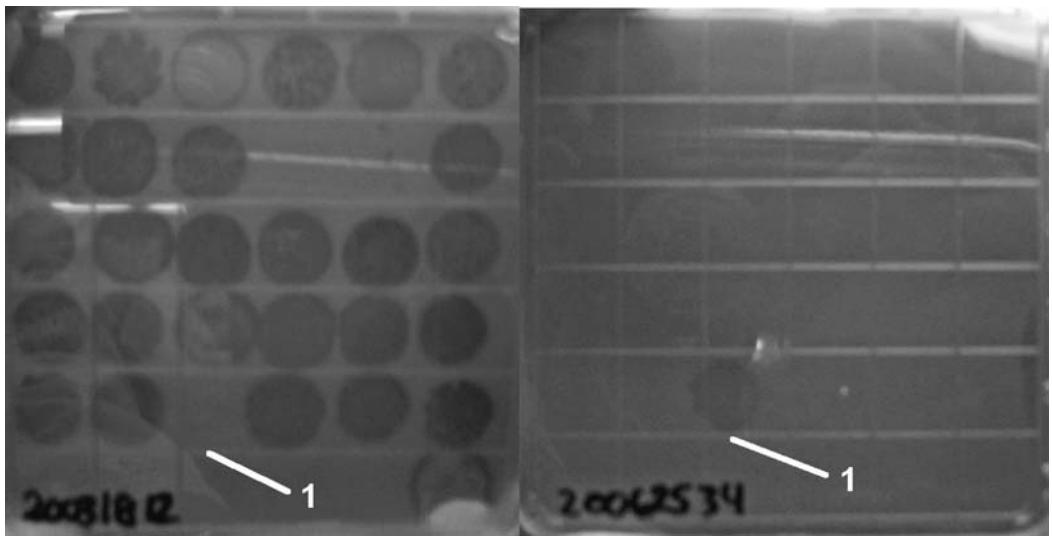


Fig. 17.4. Appearance of lysis zones on *Salmonella enterica* serovar Typhimurium PT2 (*Left*) and PT124 (*Right*). The arrowed labeled "1" points to the only phage which lyses PT124, but has no activity on PT2.

patters are available for certain serovars—i.e., *S.* Enteritidis. For comprehensive outbreak investigation, we recommend the use of multiple sub-typing techniques employed in conjunction with epidemiological information regarding human exposure to the source of contamination *(9)*.

4 Notes

1. The assumption here is that the user will be typing *Salmonella* or *E. coli*. Phage broth has the same biochemical composition as phage agar except that the agar is omitted. For other bacteria, use the medium recommended for host propagation or phage titration. Two good sources for media formulations are the American Type Culture Collection (ATCC) at http://www.atcc.org/Home.cfm, or the Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ) at http://www.dsmz.de/. If the phages require divalent cations for adsorption or replication, the medium should be appropriately supplemented.
2. It is important that the drop is large enough that the needle tips of the syringes do not touch the surface of the agar. This is to avoid contamination of the phage syringes by *Salmonella* bacteria as well as between-sample contamination of the bacterial lawn.

3. The phage agar and broth pH after sterilization in the autoclave must be adjusted and maintained at 6.8, minor deviation from the prescribed pH may adversely affect the outcome.
4. If the plates are not dried enough, phage drops will spread or run into each other and if plates are over dried then the lawn or bacterial growth will be uneven on the plate surface, making accurate lytic reaction reading impossible.

References

1. Gürtler, V. and B.C. Mayall. 2001. Genomic approaches to typing, taxonomy and evolution of bacterial isolates. International Journal of Systematic and Evolutionary Microbiology *51*:3–16.
2. Vandamme, P., P. Bots, P. Gillis, P. de Vos, K. Kersters, and J. Swings. 1996. Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiological Reviews *60*: 407–438.
3. Tenover, F.C., R.D. Arbeit, R.V. Goering, P.A. Mickelsen, B.E. Murray, D.H. Persing, and B. Swaminathan. 1995. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. Journal of Clinical Microbiology *33*:2233–2239.
4. Lindstedt, B.-A., E. Heir, T. Vardund, K.K. Melby, and G. Kapperud. 2000. Comparative fingerprinting analysis of *Campylobacter jejuni* subsp. *jejuni* strains by amplified-fragment length polymorphism genotyping. Journal of Clinical Microbiology *38*:3379–3387.
5. Gevers, D., G. Huys, and J. Swings. 2007. Applicability of rep-PCR fingerprinting for identification of *Lactobacillus* species. FEMS Microbiology Letters *205*:31–36.
6. Isaacs, S., J. Aramini, B. Ciebin, J.A. Farrar, R. Ahmed, D. Middleton, A.U. Chandran, L.J. Harris, et al. 2005. An international outbreak of salmonellosis associated with raw almonds contaminated with a rare phage type of *Salmonella enteritidis*. Journal of Food Protection *68*:191–198.
7. Woodward, D.L., C.G. Clark, R.A. Caldeira, R. Ahmed, G. Soule, L. Bryden, H. Tabor, P. Melito, et al. 2005. Identification and characterization of *Shigella boydii* 20 serovar nov., a new and emerging *Shigella* serotype. Journal of Medical Microbiology *54*:741–748.
8. McLaughlin, J., L.J. Castrodale, M.J. Gardner, R. Ahmed, and B.D. Gessner. 2006. Outbreak of multidrug resistant *Salmonella* Typhimurium associated with ground pork served at a school potluck. Journal of Food Protection *69*:666–670.
9. Ahmed, R., P. Sankar-Mistry, S. Jackson, H.-W. Ackermann, and S.S. Kasatiya. 1995. *Bacillus cereus* phage typing as an epidemiological tool in outbreaks of food poisoning. Journal of Clinical Microbiology *33*: 636–640.
10. Callow, B.R. 1922. Bacteriophage phenomena with *Staphylococcus aureus*. Journal of Infectious Diseases *30*:643–640.
11. Wentworth, B.B. 1963. Bacteriophage typing of the staphylococci. Bacteriological Reviews *27*:253–272.
12. Craigie, J. and C.H. Yen. 1938. The demonstration of types of *B. typhosus* by means of preparations of type II Vi phage. I. Principles and technique. Canadian Journal of Public Health *29*:448–484.
13. Craigie, J. and C.H. Yen. 1938. The demonstration of types of *B. typhosus* by means of preparations of type II Vi phage. II. The stability and epidemiological significance of V form types of *B. typhosus*. Canadian Journal of Public Health *29*:484–496.
14. Anderson, E.S. and R.E. Williams. 1956. Bacteriophage typing of enteric pathogens and staphylococci and its use in epidemiology. Journal of Clinical Pathology *9*:94–127.
15. Fisk, R.T. 1942. Studies on staphylococci. I. Occurence of bacteriophage carriers among strains of *Staphylococcus aureus*. Journal of Infectious Diseases *71*:153–160.
16. Yousten, A.A., B.H. de, J. Hedrick, D.V. Cosmao, and P. Myers. 1980. Comparison between bacteriophage typing and serotyping for the differentiation of *Bacillus sphaericus* strains. Annales de Microbiologie *131B*: 297–308.
17. Rabkin, C.S., W.R. Jarvis, and W.J. Martone. 1987. Current status of *Pseudomonas cepacia* typing systems. European Journal of Epidemiology *3*:343–346.
18. Grajewski, B.A., J.W. Kusek, and H.M. Gelfand. 1985. Development of a bacteriophage typing system for *Campylobacter jejuni* and *Campylobacter coli*. Journal of Clinical Microbiology *22*:13–18.

19. Hiramune, T. and R. Yanagawa. 1969. *Corynebacterium renale*, Phage-types and their epidemiological significance. Japanese Journal of Veterinary Research *17*:25–31.
20. Dei, R. and R. Dei. 1989. Observations on phage-typing of *Clostridium difficile*: preliminary evaluation of a phage panel. European Journal of Epidemiology *5*:351–354.
21. Prevot, A.R. and H. Thouvenot. 1961. Attempted phage-typing of anaerobic Corynebacteria. Annales de l'Institut Pasteur *101*:966–970.
22. Weischer, M., H.J. Kolmos, M.E. Kaufmann, and V.T. Rosdahl. 1993. Biotyping, phage typing, and O-serotyping of clinical isolates of *Enterobacter cloacae*. APMIS *101*: 838–844.
23. Ahmed, R., C. Bopp, A. Borczyk, and S. Kasatiya. 1987. Phage-typing scheme for *Escherichia coli* O157:H7. Journal of Infectious Diseases *155*:806–809.
24. Frost, J.A., H.R. Smith, J.A. Willshaw, S.M. Scotland, R.J. Gross, and B. Rowe. 1989. Phage-typing of Vero-cytotoxin (VT) producing *Escherichia coli* O157 isolated in the United Kingdom. Epidemiology & Infection *103*:73–81.
25. Nicolle, P., L. Le Minor, R. Buttiaux, and P. Ducrest. 1952. Phage typing of *Escherichia coli* isolated from cases of infantile gastroenteritis. II. Relative frequency of types in different areas and the epidemiological value of the method. Bulletin de l Academie Nationale de Medecine *136*:483–485.
26. Parisi, J.T., J.C. Russell, and R.J. Merlo. 1969. Bacteriophage typing as an epidemiological tool for urinary *Escherichia coli*. Applied Microbiology *17*:721–725.
27. Rocourt, J., A. Schrettenbrunner, and H.P. Seeliger. 1982. Isolation of bacteriophages from *Listeria monocytogenes* Serovar 5 and *Listeria innocua*. Zentralblatt fur Bakteriologie, Mikrobiologie und Hygiene – 1 – Abt – Originale A, Medizinische Mikrobiologie, Infektionskrankheiten und Parasitologie *251*: 505–511.
28. Ikeda, Y., H. Saito, K. Miura, J. Takagi, H. Aoki, Y. Ikeda, H. Saito, K.i. Miura, et al. 2004. DNA base composition, susceptibility to bacteriophages and interspecific transformation as criteria for classification in the genus *Bacillus*. Journal of General & Applied Microbiology *50*:353–362.
29. Sechter, I., F. Mestre, and D.S. Hansen. 2000. Twenty-three years of *Klebsiella* phage typing: a review of phage typing of 12 clusters of nosocomial infections, and a comparison of phage typing with K serotyping. Clinical Microbiology & Infection *6*:233–238.
30. Nielsen, J.P. and V.T. Rosdahl. 1990. Development and epidemiological applications of a bacteriophage typing system for typing *Pasteurella multocida*. Journal of Clinical Microbiology *28*:103–107.
31. Shvidenko, I.G. and I.G. Shvidenko. 1986. Comparative evaluation of different methods for typing bacteria in the genus *Proteus*. Zhurnal Mikrobiologii, Epidemiologii i Immunobiologii 21–24.
32. Vieu, J.F. and J.F. Vieu. 1958. Preliminary note on phage typing of *Proteus hauseri*. Zentralblatt fur Bakteriologie, Parasitenkunde, Infektionskrankheiten und Hygiene - 1 - Abt - Medizinisch-Hygienische Bakteriologie, Virusforschung und Parasitologie - Originale *171*:612–615.
33. Postic, B., M. Finland, B. Postic, and M. Finland. 1961. Observations on bacteriophage typing of *Pseudomonas aeruginosa*. Journal of Clinical Investigation *40*:2064–2075.
34. Pavlatou, M.P., E. Hassikou-Kaklamani, M.P. Pavlatou, and E. Hassikou-Kaklamani. 1961. Research on phage-typing of pyocyanic bacilli. Annales de l'Institut Pasteur *101*:914–927.
35. Demczuk, W., G. Soule, C. Clark, H.W. Ackermann, R. Easy, R. Khakhria, F. Rodgers, and R. Ahmed. 2003. Phage-based typing scheme for *Salmonella enterica* serovar Heidelberg, a causative agent of food poisonings in Canada. Journal of Clinical Microbiology *41*: 4279–4284.
36. Callow, B.R. 1959. A new phage-typing scheme for *Salmonella typhi-murium*. Journal of Hygiene *57*:346–359.
37. Anderson, E.S., L.R. Ward, M.J. De Saxe, and J.D.H. De Sa. 1977. Bacteriophage-typing designations of *Salmonella* Typhimurium. Journal of Hygiene *78*:297–300.
38. Felix, A. and B.R. Callow. 1943. Typing of paratyphoid B bacilli by means of Vi bacteriophage. British Medical Journal *ii*: 127–130.
39. Hamilton, R.L. and W.J. Brown. 1972. Bacteriophage typing of clinically isolated *Serratia marcescens*. Applied Microbiology *24*: 899–906.
40. Hammarstrom, E. 1949. Phage typing of *Shigella sonnei*. Acta Medica Scandinavica *133 (Suppl. 223)*:1–132.
41. Hill, A.W. and C.A. Brady. 1989. A note on the isolation and propagation of lytic phages from *Streptococcus uberis* and their potential for strain typing. Journal of Applied Bacteriology *67*:425–431.
42. Kuhnen, E., F. Richter, K. Richter, and L. Andries. 1988. Establishment of a typing system for group D streptococci. Zentralblatt Fur Bakteriologie, Mikrobiologie, Und Hygiene –

Series A, Medical Microbiology, Infectious Diseases, Virology, Parasitology *267*: 322–330.
43. Birzu, A., P. Plecea, H. Aizicovici, and V. Moroanu. 1966. Bacteriophage typing of enterococci isolated from food and from cases of food poisoning. Archives Roumaines de Pathologie Experimentales et de Microbiologie *25*:245–252.
44. Saldanha, F.L., B.C. Gandhi, S.N. Sayyid and R.S. Bhave 1965. Recent observations on sero-types and phage-types of *V. cholerae* in Maharashtra. Indian Journal of Medical Research *53*:926–933.
45. Hoi, L., I. Dalsgaard, A. DePaola, R.J. Siebeling, and A. Dalsgaard. 1998. Heterogeneity among isolates of *Vibrio vulnificus* recovered from eels (*Anguilla anguilla*) in Denmark. Applied & Environmental Microbiology *64*:4676–4682.
46. Nicolle, P., H. Mollaret, Y. Hamon, J.F. Vieu, J. Brault, and G. Brault. 1967. Lysogenic, bacteriocinogenic and phage-typing study of species *Yersinia enterocolitica*. Annales de l'Institut Pasteur *112*:86–92.
47. Adlakha, S., K.B. Sharma, K. Prakash, S. Adlakha, K.B. Sharma, and K. Prakash. 1986. A new phage typing scheme for strains of *Salmonella typhimurium* isolated in India. Indian Journal of Medical Research *84*: 14–19.
48. Schmieger, H. 1999. Molecular survey of the *Salmonella* phage typing system of Anderson. Journal of Bacteriology *181*: 1630–1635.
49. Anderson, E.S. 1964. The phage typing of *Salmonella* other than *S*. Typhi, p. 89–109. *In* E. Van Oye (Ed.), The World Problem of Salmonellosis. Dr.W.Junk Publishers, The Hague.
50. Felix, A. 1956. Phage typing of *Salmonella typhimurium*: its place in epidemiological and epizootiological investigations. Journal of General Microbiology *14*:208–222.
51. Ibrahim, A.A.E. 1969. Bacteriophage typing of *Salmonella* I. Isolation and host range study of bacteriophage. Applied Microbiology *18*:444–447.
52. Lilleengen, K. 1948. Typing *Salmonella typhimurium* by means of bacteriophage. Acta Pathologica Microbiologica Scandanavica Supplimentum *77*:11–128.
53. Wilson, V.R., G.J. Hermann, and A. Balows. 1971. Preliminary report on a new system for typing *Salmonella typhimurium* in the United States. Applied Microbiology *21*: 774–776.
54. Rabsch, W., R.A. Helm, and A. Eisenstark. 2004. Diversity of phage types among archived cultures of the Demerec collection of *Salmonella enterica* serovar Typhimurium strains. Applied & Environmental Microbiology *70*:664–669.
55. Pruneda, R.C. and J.J.I. Farmer. 1977. Phage typing of *Shigella sonnei*. Journal of Clinical Microbiology *5*:66–74.

Chapter 18

A Genetic Screen to Identify Bacteriophage Lysins

Raymond Schuch, Vincent A. Fischetti, and Daniel C. Nelson

Abstract

Lysins are phage-encoded, peptidoglycan (cell wall) hydrolases that accumulate in the bacterial cytoplasm during a lytic infection cycle. Late during infection, the lysins undergo holin-mediated translocation across the inner membrane into the peptidoglycan matrix where they cleave cell wall covalent bonds required for wall stability and allow bacterial lysis and progeny phage release. This potent hydrolytic activity is now the foundation of a powerful genetic-based screening process for the identification and analysis of phage lysin proteins. Here, we describe a method for identifying a lysin, PlyG, from a bacteriophage that specifically infects the Gram-positive organism *Bacillus anthracis*; however, the techniques described can be adapted to clone, express, and analyze lysins from any phage infecting Gram-positive bacteria or possibly even Gram-negative bacteria.

Key words: Lysin, hydrolase, cell wall, peptidoglycan, lysozyme, Gram-positive, antimicrobial, diagnostic, expression library.

1 Introduction

A lysin-based lytic mechanism *(1)* is widespread amongst phages infecting both Gram-positive and Gram-negative bacteria. Indeed, the lysins likely represent one of the most successful bacteriolytic agents used in nature *(2)*. A major distinguishing feature of the lysin family of proteins concerns their modular design, which can vary dramatically depending on whether a lysin targets the thick, surface exposed cell wall of Gram-positive bacteria, or the very thin peptidoglycan of Gram-negative organisms that lies subjacent to and is protected by the external outer membrane. As such, lysins active against Gram-negative bacteria usually consist of a single 15–20 kDa catalytic domain, while those active against Gram-positive bacteria are 25–100 kDa and possess one

Martha R. J. Clokie, Andrew M. Kropinski (eds.), *Bacteriophages: Methods and Protocols, Volume 2: Molecular and Applied Aspects, vol. 502*, © 2009 Humana Press, a part of Springer Science+Business Media
DOI 10.1007/978-1-60327-565-1_18 Springerprotocols.com

or more distinct catalytic domains (multi-domain protein) fused to a *c*ell *w*all-*b*inding domain (CBD). The CBDs of Gram-positive lysins are believed to recognize strain- or species-specific carbohydrate structures that decorate the surface of the Gram-positive cell wall *(3)*, thus exerting a level of specificity over the binding of such lysins. A potent lytic activity, combined with an often extreme binding specificity, distinguishes the lysins of Gram-positive microorganisms.

Unlike with Gram-negative peptidoglycan, the surface-exposed nature of Gram-positive peptidoglycan renders it vulnerable to hydrolysis by exogenously applied lysins. We have now demonstrated this "lysis from without" using an array of different lysins specifically active against a broad range of Gram-positive bacteria *(4)*. Based on these findings, our laboratory and others are actively pursuing development of lysins as novel therapeutic agents (or alternatives to conventional antibiotics) active against Gram-positive pathogens. Additionally, we are seeking to exploit the binding capacity of these enzymes as the basis for sensitive diagnostic methods. The genetic and biochemical methods developed for these studies are presented.

2 Materials

2.1 Phage, Plasmids, and Bacteria

1. The γ phage, pBAD24 *(5)*, and *Bacillus cereus* strain RSVF1 are part of The Rockefeller University Collection.
2. *E. coli* XL2-Blue ultracompetent cells (Stratagene, La Jolla, CA; http://www.stratagene.com/)

2.2 Equipment

1. RunOne electrophoresis multicasting system (EmbiTec, San Diego, CA; http://www.embitec.com/).
2. Model C24 incubator shaker (New Brunswick Scientific, Edison, NJ; http://www.nbsc.com/Main.asp).
3. SpectraMax Plus spectrophotometer (Molecular Devices, Sunnyvale, CA; http://www.moleculardevices.com/).
4. Model 5810R tabletop centrifuge (Eppendorf).
5. Orbital Shaker (Bellco Biotechnology, Inc., Vineland, NJ; http://www.bellcoglass.com/).
6. EchoTherm Model IC20 (Torrey Pines Scientific, San Marcos, CA; http://www.torreypinesscientific.com/) for incubations at 16 °C and 65 °C.
7. UV transluminator (Alpha Innotech Corp., San Leandro, CA; http://www.alphainnotech.com/).
8. Dual chamber water bath (Precision).

2.3 Supplies

1. Lambda Maxi Kit (Qiagen, Inc., Valencia, CA; http://www1.qiagen.com/).

2. All DNA modifying enzymes and corresponding buffers were purchased from New England Biolabs (Ipswich, MA; http://www.neb.com/nebecomm/default.asp).
3. 1.5 ml Eppendorf tubes.
4. Phenol, chloroform, NH_4OAc, ethanol, ethidium bromide, NaOH, Tris, L-arabinose, phosphate buffered saline (PBS), LB media, and ampicillin were all from Sigma-Aldrich (St. Louis, MO; http://www.sigmaaldrich.com/).
5. 1 kb DNA ladder (Invitrogen Corp., Carlsbad, CA; http://www.invitrogen.com/).
6. TAE buffer.
7. Agarose (Cambrex Corp.; http://www.cambrex.com/default.asp).
8. NucleoSpin DNA extraction columns and NucleoSpin Plasmid Kit (BD Biosciences, San Jose, CA; http://www.bdbiosciences.com/).
9. 150 × 15 mm polystyrene and 150 mm glass Petri dish (Fisher, Inc.).
10. Replica transfer apparatus and velvet squares (Cat. No. 11DOTM001; MP Biochemicals; Solon, OH; http://www.mpbio.com/).
11. 10 ml round bottom culture tube (Sarstedt).
12. Brain Heart Infusion media (Difco).
13. 96-well microtiter plate (Costar).
14. 125 ml and 2 L Ehrlenmeyer flasks (VWR).
15. NucleoSpin DNA extraction columns (Clontech—a Takara Bio Company, Mountain View, CA; http://www.clontech.com/clontech/).

3 Methods

The potent bacteriolytic activity of phage lysins provides the basis by which their coding sequences may be identified *(6)*. Toward this end, we have developed an efficient means to detect lysin activity by screening induced plasmid expression libraries for agents that lyse live bacterial cells. We describe here one such activity screen for identification of a lysin, PlyG, from the purified bacteriophage, γ. The γ phage is a diagnostic tool used in clinical laboratories to identify the Gram-positive pathogen *Bacillus anthracis*, although it also infects some highly related *B. cereus* isolates like RSVF1 *(7)*.

Following our description of lysin identification, we detail the processes by which we define and quantify lysin activity. Owing to the conserved, repeating structure of bacterial peptidoglycan, lysin catalytic domains are limited to one of

four enzymatic activities: *N*-acetylglucosaminidase, *N*-acetylmuramidase (lysozyme), *N*-acetylmuramoyl-L-alanine amidase, or endopeptidase. Unfortunately, because the catalytic domains from Gram-positive phage lysins are not enzymatically active *in vitro*, there is no assay involving simple chromogenic or flurogenic substrates that mimic target cell wall bonds and allow kinetic measurements or even simple activity quantitation. A higher order structure, perhaps one containing the CBD carbohydrate-binding epitope in association with peptidoglycan, is apparently necessary for activity. Thus, the lysins require either purified cell walls, or whole cells as substrates for *in vitro* analyses. As result, we have adapted a turbidimetric assay *(8)* to titer lysin activity and define a standard unit. This screening method is based on observing zones of clearing on lawns of bacterial cells and produces results which are reproducible and allow direct comparison of activity between lysins active on different species.

It is important to note that the details of lysin identification, purification, and analysis will require slight, yet significant, alterations based on variables like the intended lysin source (i.e., is it encoded within a lytic phage, bacterial genome (prophage), or complex environmental sample) and activity target (i.e., Gram-positive vs. Gram-negative organism, fast or slow growing, etc.). Variations that may be incorporated into a lysin screen are described in the **Notes**. Otherwise, the methods described here were specifically developed to identify the PlyG lysin encoded within the purified genomic DNA of γ phage, although this method has also identified lysins active against *Bacillus cereus, B. thuringiensis, Enterococcus faecalis, E. faecium, Listeria monocytogenes, Streptococcus pyogenes, S. pneumoniae*, and *S. agalactiae (7, 9–13)*.

3.1 Construction of a Plasmid Expression Library Encoding the Partially Digested γ Phage Genome

1. Prepare total genomic DNA from 250 ml of high-titer ($\sim 1 \times 10^9$ pfu ml^{-1}) γ phage lysate using the Qiagen Lambda Maxi Kit (*see* **Note 1**). Resuspend purified DNA in distilled water at a final concentration of 500 ng μl^{-1}.
2. Prepare five 1.5 ml Eppendorf tubes, each containing 5 μg of phage DNA in 50 μl 1X New England Biolabs (NEB) Buffer 2. Add the NEB restriction endonuclease *Tsp*509I (*see* **Note 2**) to final enzyme unit amounts of 10, 5, 2.5, 1.0 and 0.5, respectively. Gently suspended the enzyme and incubate tubes for 5 min at 65 °C.
3. Since *Tsp509I* cannot be heat inactivated, perform the following: immediately add 50 μl phenol:chloroform (1:1 mixture) to each tube, vortex for 10 s, and centrifuge for 5 min at 4 °C in a table-top microcentrifuge (maximum speed). Recover the upper DNA phase, add 50 μl chloroform, and again vortex and centrifuge. To the resultant upper phase, add 50 μl of

5 M NH_4OAc, mix by inversion, add 250 µl 100% ethanol, and mix again. Centrifuge for 15 min at 4 °C, wash each pellet with 70% ethanol, and resuspend in 15 µl of dH_2O.

4. Prepare a 0.4 mm thick, 1% agarose gel. Add gel-loading buffer to each tube and load each (as well as a lane for 1 Kb DNA ladder) into gel with 1X TAE running buffer. The bromophenol blue dye front should be run to the bottom of the gel.
5. Stain gel for 15 min in 1X TAE containing 0.5 µg ml^{-1} ethidium bromide and observe on a UV transilluminator, keeping exposure time to a bare minimum. With a clean razor blade, carefully excise gel segments containing partially digested DNA fragments in only the range of 500–2,000 bp (*see* **Note 3**).
6. Recover DNA from agarose slices using 4–8 NucleoSpin DNA Extraction Columns and elute each in a final volume of 50 µl distilled water (dH_2O). Pool samples and run 10 µl on a 1% agarose gel for quantitation of DNA recovery—a DNA smear should be seen in the 500–2,000 bp range.
7. Set up three 15 µl ligations in 0.2 ml tubes containing 1X ligation buffer, 10 units of T4 DNA ligase and ~10 ng of plasmid pBAD24 (previously linearized with *Eco*RI and dephosphorylated with antarctic phosphatase) and either 50, 100, or 200 ng, respectively, of *Tsp*509I-digested phage DNA. Set up a fourth ligation, in which all phage DNA is omitted, is established as a self-ligation control.
8. After overnight incubation at 16 °C, transform 2 µl of each ligation into 100 µl XL2-Blue ultracompetent *E. coli* following the manufacturer's protocol (Stratagene, Inc.).
9. Plate cells on LB agar supplemented with 100 µg ml^{-1} ampicillin, at a density yielding ~300 distinct colonies per 150 × 15 mm polystyrene Petri dish after overnight incubation at 37 °C. Ultimately, 10 plates, each containing ~300 colonies, should be obtained (*see* **Note 4**).
10. The quality of the γ phage expression library, with respect to insert size and diversity, must be assessed. First, isolate total DNA from 25 random transformants by suspending each colony in 25 µl of 0.5 M NaOH, and adding 25 µl 1 M Tris (pH 8.0), 450 µl dH_2O, and mixing vigorously. Next, use 1.0 µl DNA samples as template in PCRs with primers that flank the pBAD24 MCS (*see* **Note 5**). Run 7.5 µl of each reaction on a 1% agarose gel to assess the range of insert sizes. We observed that 24 of 25 transformants had γ phage inserts (from 0.3 to 2.5 kb in size) with no bias toward any particular size as expected for a truly random library (*see* **Note 6**).

3.2 Screening the γ phage Library for Lysin Activity

1. The γ phage expression library consists of ∼3,000 *E. coli* transformants on 10 150 × 15 mm LB plates containing ampicillin.
2. Using 150 mm velvet squares and a replica transfer apparatus, replica plate library onto 10 150 mm glass Petri dishes, each consisting of 90 ml LB agar with 100 μg μl^{-1} ampicillin and 0.2% L-arabinose (*see* **Note 7**).
3. Mark the master and replica plates with an alcohol-resistant marker at such a position to allow accurate alignment of plate pairs when recovering positive clones.
4. Incubate plates overnight at 37 °C for the induced library to grow in.
5. On the day prior to the activity screen, set up a 5 ml LB liquid culture with *B. cereus* strain RSVF1 (*see* **Note 8**) and incubate overnight at 30 °C shaking at 150 rpm.
6. On the day of the activity screen, have available: 55 °C LB soft agar (0.75% agar) and a set of 10 10 ml round-bottom culture tubes each containing 100 μl of *B. cereus* RSVF1 overnight culture and kept at 37 °C.
7. Permeabilize induced *E. coli* expression clones with chloroform vapors as follows: in a chemical fume hood, add 10 ml chloroform to the inverted lid of each glass plate, then invert *E. coli* clones over the chloroform and incubate for 10 min, then place the clones face up for 10 min to allow chloroform evaporation (*see* **Note 9**).
8. Add 7.5 ml LB soft agar to an RSVF1-containing tube and pour contents over the clones on the glass plate, rapidly rotating to allow overlay of the entire surface (*see* **Note 10**). Repeat for each glass plate.
9. Incubate plates at room temperature for 4 h (the RSVF1 lawn may become visible as a very faint haze at this point). Place at 4 °C overnight and return to room temperature the following day to allow the RSVF1 to grow in.
10. Lysin-encoding clones are distinguished by the appearance of distinct RSVF1-clearing zones in the overlay surrounding such clones (*see* **Fig. 18.1**). If no zones become apparent after incubation at room temperature, incubate overnight at 4 °C again, then return to room temperature and continue search for clearing zones (*see* **Note 11** for alternative methods to screen for lysins). We detected 52 positive clones in our screen (1.7% of the library). In similar screens for other *Siphoviridae* lysins, we observed rates between 0.1 and 2%.
11. Recover master *E. coli* colonies corresponding to positive clones by aligning glass plate-master polystyrene plate pairs. Streak to single colonies on LB agar plates containing 100 μg μl^{-1} ampicillin and incubate overnight at 37 °C.

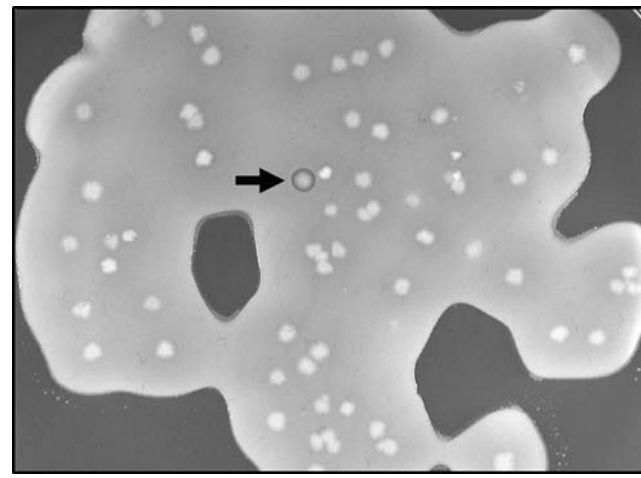


Fig. 18.1. Lytic activity screen for the identification of γ phage lysin, PlyG. The induced and permeabilized *E. coli* expression library is overlaid with agar containing *B. cereus* strain RSVF1. Owing to the presence of ampicillin in the LB agar, the ampicillin-sensitive RSVF1 grows only near *E. coli* colonies expressing plasmid-encoded β-lactamase. A clearing zone in the RSVF1 overlay, surrounding one of the library clones (indicated by an *arrow*), identifies the presence of a cloned lysin.

12. Restreak each clone on LB agar with ampicillin and inoculate 5 ml LB liquid with ampicillin and incubate overnight at 37 °C shaking at 200 rpm.
13. Use the LB agar plate cultures to generate frozen stocks (*see* **Note 12**).
14. Use the LB liquid cultures to prepare plasmid using the NucleoSpin Plasmid Kit. Sequence resulting plasmids with both BAD1 and BAD3 primers (*see* **Note 5**).
15. Assemble DNA sequences and examine *in silico* to determine whether a lysin has been identified (*see* **Note 13**).

3.3 Expression of the PlyG lysin

1. To create a PlyG expression strain, clone the entire 702 bp *plyG* ORF (include no flanking γ phage sequence) into plasmid expression vector pBAD24 and transform *E. coli* strain XL1-Blue (*see* **Note 14**).
2. Inoculate XL1-Blue/pBAD24::*plyG* into 15 ml LB liquid culture with 100 μg ml^{-1} ampicillin (in a 125 ml Ehrlenmeyer flask) and shake at 30 °C overnight at 150 rpm.
3. Dilute overnight culture 1:100 into 1 L of LB liquid with ampicillin (in a 2 L Ehrlenmeyer flask) and shake at 225 rpm for 3 h at 37 °C.
4. Add the L-arabinose inducer to a final concentration of 0.2% and continue growth at 30 °C overnight (~12 h) (*see* **Note 15**).

5. Wash culture by spinning down cells at 4,000 rpm in an Eppendorf tabletop centrifuge and then adding a 1X volume of 1X PBS to the pellet, remove the supernatant, and finally resuspend in 50 ml 1X PBS.
6. For bacterial lysis, add chloroform to a final concentration of 20%, vortex for 1 min, gently agitate in an orbital shaker for 1 h at 4 °C, and vortex again for 1 min (*see* **Note 16**).
7. Pellet bacterial debris for 15 min at 4 °C and 4,000 rpm in an Eppendorf tabletop centrifuge and carefully remove supernatant.
8. The supernatant contains crude PlyG, which can be assayed for activity according to **Section 3.4** (*see* **Note 17**).

3.4 Quantifying Lysin Activity

1. Grow *B. cereus* strain RSVF1 overnight in 10 ml Brain Heart Infusion (BHI) broth at 30 °C shaking at 150 rpm.
2. The next morning, start a 50 ml culture of RSVF1 from a 1:100 dilution of the overnight in fresh BHI media and incubate as above.
3. At exactly 3 h of incubation for this culture (*see* **Note 18**), harvest the cells by centrifugation in 50 ml Falcon tubes for 15 min at 4 °C and 4,000 rpm in an Eppendorf tabletop centrifuge.
4. Resuspend the pellet and wash 1X in 10 ml of phosphate-buffered saline (PBS) (*see* **Note 19**).
5. Resuspend the pellet and adjust the final OD_{600} to 1.0 by the addition of PBS.
6. Prepare PlyG for assay by diluting PlyG 1:10 in PBS and making additional twofold dilutions in PBS (i.e., final dilutions will be 1:10, 1:20, 1:40, 1:80, etc.).
7. Pre-warm a 96-well microtiter plate and Molecular Devices spectrophotometer to 37 °C (*see* **Note 20**).
8. Add 100 µl of the PlyG dilutions in duplicate to the 96-well plate. Include a "no PlyG" control with 100 µl PBS only.
9. At time zero, add 100 µl aliquots of the *B. cereus* strain RSVF1 to each well using a 12-channel micropipette.
10. Agitate for 5 sec and take an OD_{600} endpoint reading. Since the starting OD_{600} was adjusted to 1.0, and these cells were mixed 1:1 with PlyG/PBS, all wells should now have a reading of ~ 0.5.
11. Allow the door of the spectrophotometer to close, incubate at 37 °C with continual agitation, and take another OD_{600} endpoint reading at time = 15 min.
12. Measurement of lysin activity is based on turbidimetric determination of cell lysis. The 15 min endpoint readings should form a range from 0.5 (for controls or very dilute PlyG wells) to 0.1 for concentrated PlyG wells. Thus, we *define* the standard lysin "unit" as the highest dilution that produces a

half-drop in OD in the 15 min assay. For example, if an 80-fold dilution of PlyG produced a drop in OD_{600} of *B. cereus* strain RSVF1 from 0.5 to 0.25 in 15 min, then we say that PlyG has a titer (or activity) of 80 U/ml for this strain (*see* **Note 21**).

13. It is our experience that most lysins, when purified to homogeneity, will have a specific activity of ~1 U per μg protein, although some very active lysins may have a specific activity 100 or 1,000 times higher.

4 Notes

1. For sufficient DNA yield, starting phage titers must be $>1 \times 10^8$ pfu ml^{-1}. Additionally, the Lambda Maxi Kit is effective only for long-tailed phage of the families *Siphoviridae* and *Myoviridae*. For *Podoviridae* (short tails) and *Tectiviridae* (no tails), we incorporated a higher *g*-force spin (ultracentriguation at 35,000 rpm for 3 h) after the PEG precipitation of phage or purification by CsCl gradient. See **Chapter 34** for additional protocols of phage purification and Chapter 22 for protocols on DNA extraction.
2. *Tsp*509I was chosen for its ability to completely degrade the γ genome (no obvious fragments >500 bp) after overnight digestion. Additionally, it generates cohesive ends (/AATT), compatible with *Eco*RI in the pBAD24 MCS. We generally use *Tsp*509I for phage of Gram-positive bacteria. For phage of Gram-negative bacteria, *Hae*III (GG/CC) is often appropriate for cloning into the *Sma*I site of pBAD24.
3. With an effective partial digestion, the highest and lowest enzyme unit lanes should be completely degraded (a DNA smear smaller than 500 bp) and uncut (a single high-molecular weight band), respectively. With intermediate concentrations, a smear of progressively lower-molecular weight should be observed with increasing enzyme concentration. The lane/s in which a majority of products are between 500 and 2,000 bp is appropriate for excision. It may be necessary to repeat the reactions, modifying only the amount of enzyme, in order to get a proper partial digestion pattern.
4. To obtain 300 colonies per plate, trials may be required in which the amount of final transformation mix plated is varied. We first plate 50 and 500 μl of the 1 ml mixture and then, based on resulting colony counts, we repeat the transformation and plate appropriately on 10 plates.
5. For sequence analysis of pBAD24 inserts, we use the following primers: BAD1, 5′-CTACTGTTTCTCCATACC-3′ and BAD3, 5′-GCAGTTCCCTACTCTCGC-3′. PCRs use

the following conditions: 30 cycles of 95 °C at 30 s, 50 °C at 30 s, and 72 °C at 2 min.

6. For less efficient libraries, we simply screen a greater number of transformants. The number of insert-bearing transformants that must be screened to identify a lysin gene is predicted with the following equation: 1/(size of the coding region of interest/(size of phage genome × 2)), where the size of the coding region is 700 bp (average size for *Bacillus spp.* phage lysins), the size of the genome is estimated at 40,000 bp (roughly average size for *Siphovidae* phage of *Bacillus spp.*). The factor 2 accounts for two possible insert orientations in the vector. This equation predicts that we must screen 114 transformants to identify *plyG* (we screened 3,000 to be cautious and found 52 *plyG* positive clones).
7. After pouring, let glass plates dry completely at room temperature prior to replica plating. Once poured, these may be stored at 4 °C for up to no more than 1 month.
8. RSVF1 is *B. cereus* strain ATCC 4342.
9. While the library is being permeabilized, chloroform has a tendency to pool at plate edges or be drawn up between the lid and base. To avoid this, occasionally agitate plates to redistribute chloroform over the surface of the inverted lid. During the chloroform evaporation step, residual chloroform from each lid is discarded, and the lids are washed quickly with soap and water, rinsed and dried.
10. It is extremely difficult to distribute agar evenly over the entire surface if plates are not at room temperature. To ensure that the overlay process occurs rapidly, we perform this step next to a 55 °C water bath containing the LB soft agar and a 37 °C water bath with the RSVF1 tubes.
11. Although we describe a lysin activity screen, several alternate methods for lysin identification have been developed and used in our lab. One method is the holin activity screen *(7)*: perform this exactly as described for the lysin screen, with the exception that the clones which grow well in non-induced conditions (i.e., master plates) but poorly in induced conditions (i.e., glass plates) can be used to identify a holin gene, which is usually encoded immediately upstream of a lysin. Additional methods involve, (**a**) using known lysin DNA sequences in BLAST- or PCR-based screens (using primers directed against well-conserved catalytic domains) of bacterial genomes, (**b**) direct purification of lysins from phage lysates followed by protein sequence analysis, and (**c**) SDS-PAGE separation of phage proteins followed by gel overlays containing dense bacterial cell wall material; subsequent clearing zones identify the phage lysin bands that may then be sequenced, and **d**) total phage genome sequencing.

12. To make a frozen cell stock, freshly plated bacteria are scraped up from the plate surface and suspended in 1.5 ml cryovials with 1 ml of LB with 15% glycerol. After 15 min incubation, the tubes are placed in the −70 °C freezer.
13. Raw insert sequence is subjected to BLASTX or ORF Finder (NCBI) analysis. A lysin insert should be similar to other lysins in the protein sequence database. Sequence alignments can be used to assess degrees of relatedness. Alternately, in lieu of significant primary sequence homology, searching the Pfam (protein families) database (http://www.sanger.ac.uk/Software/Pfam) is useful to identify conserved domains common to some lysins, such as the recently discovered cysteine/histidine aminopeptidase (CHAP) domain *(14)*. Generally, lysins of Gram-positive bacteria have hallmark structure: a well-conserved N-terminal catalytic domain and a poorly conserved C-terminal cell wall-binding domain.
14. For cloning of *plyG* (accession number AF536823), primers including both the first 20 (PLYG1) and last 20 (PLYG2) bases of the ORF were used, with *Eco*RI and *Hin*dIII sites incorporated into their ends, respectively. The resulting PCR product was cloned into the *Eco*RI–*Hin*dIII sites of plasmid pBAD24.
15. Expression will have to be optimized for each lysin. We have found that induction in 0.2% arabinose at 30 °C overnight is optimum for PlyG. However, some lysins are degraded by *E. coli* proteases with long inductions.
16. Any method of cell disruption is suitable (French Press, homogenizer, etc.). Chloroform is quick, inexpensive, gives adequate yields, and does not appear to harm lysins. However, chloroform is not practical for large, scale-up preparations (>10 L cultures).
17. We are not detailing the purification of PlyG, simply because the purification scheme for any given lysin is specific for that protein and is not necessarily applicable as a protocol for other lysins. This being said, most lysins are fairly stable and purify easily with anion exchange or cation exchange column chromatography depending on the isoelectric point of the protein. Tags or fusion proteins can be added to lysin genes in order to simplify purification (i.e. His tagging the protein and purifying on a nickel column). However, in our experience, tags are potentially deleterious to the enzymatic activity of lysins so our personal preference is to express untagged lysins and elucidate a custom purification scheme for each lysin.
18. The time of growth for the day culture is very important. Since lysins are cell wall hydrolases, their actions are more pronounced on mid-log cells than stationary cells, which are

more highly crosslinked. To standardize this assay, we always determine the units of lysin activity on mid-log phase cells. For *B. cereus* strain RSFV1, we have found that a 3-hr culture taken from a 1:100 dilution of an overnight is appropriate. This timing may be different for other bacterial species.

19. All assays to titer PlyG and determine units of activity take place in physiological phosphate buffered saline (PBS). While PBS may not offer the exact optimal pH or salt concentration for maximum PlyG activity, PBS is nonetheless chosen as the standard to titer all lysins so a direct comparison can be made between lysins.
20. For automation and simplicity, we utilize a 96-well plate spectrophotometer (SpectraMax Plus). This device allows us to monitor the drop in OD in real time (kinetic read), perform 96 simultaneous assays with reference subtraction, incubate samples at 37 °C, and continually agitate the 96-well plate so any observed drop in OD is not due to settling of bacterial cells. In practice, only time 0 and 15 min endpoint readings are needed to calculate the titer. Therefore, any spectrophotometer will work and assays can be carried out in cuvettes or test tubes as long as the proper proportions and dilutions are maintained and incubations are kept at 37 °C with a water bath or heated spectrophotometer.
21. Units of activity for a particular lysin will vary on different strains and/or species. Similar to the way, a host range is determined by titering a phage on different bacterial species, lysin host ranges can be determined by using identical amounts of lysin on various species and measuring the activity in terms of units. In some cases, the lysin host range and the corresponding phage host range will be identical, as is the case with PlyG and the γ phage. In other instances, the lysin will have a dramatically different host range. For example, the C_1 bacteriophage only forms plaques on Group C streptococci. However, its corresponding lysin is active against Groups A, C, and E streptococci.

Acknowledgements

This work was supported by grants from the James D. Watson Investigator Program of the New York State Office of Science, Technology, and Academic Research (NYSTAR) to D.C.N., the Northeast Biodefense Center (AI57158 NBC-Lipkin) to R.S., and the Defense Advanced Research Projects Agency (DARPA) and USPHS grant AI057472 to V.A.F.

References

1. Young, R., *Bacteriophage lysis: Mechanism and regulation*. Microbiol. Rev., 1992. **56**(3): 430–481.
2. Brussow, H. and R.W. Hendrix, *Phage genomics: Small is beautiful.* Cell, 2002. **108**: 13–16.
3. Loessner, M.J., et al., *C-terminal domains of Listeria monocytogenes bacteriophage murein hydrolases determine specific recognition and high-affinity binding to bacterial cell wall carbohydrates.* Molecular Microbiol., 2002. **44**(2): 335–349.
4. Fischetti, V.A., *Bacteriophage lytic enzymes: novel anti-infectives.* Trends Microbiol, 2005. **13**(10): 491–496.
5. Guzman, L.M., et al., *Tight regulation, modulation, and high-level expression by vectors containing the arabinose P_{BAD} promoter.* J. Bacteriol., 1995. **177**: 4121–4130.
6. Loessner, M.J., et al., *Three Bacillus cereus bacteriophage endolysins are unrelated but reveal high homology to cell wall hydrolases from different bacilli.* J. Bacteriol., 1997. **179**(9): 2845–2851.
7. Schuch, R., D. Nelson, and V.A. Fischetti, *A bacteriolytic agent that detects and kills Bacillus anthracis.* Nature, 2002. **418**: 884–889.
8. Fischetti, V.A., E.C. Gotschlich, and A.W. Bernheimer, *Purification and physical properties of group C streptococcal phage-associated lysin.* J. Exp. Med., 1971. **133**(5): 1105–1117.
9. Nelson, D., L. Loomis, and V.A. Fischetti, *Prevention and elimination of upper respiratory colonization of mice by group A streptococci using a bacteriophage lytic enzyme.* Proc. Natl. Acad. Sci. U.S.A., 2001. **98**: 4107–4112.
10. Yoong, P., et al., *Identification of a broadly active phage lytic enzyme with lethal activity against antibiotic-resistant Enterococcus faecalis and Enterococcus faecium.* J. Bacteriol., 2004. **186**: 4808–4812.
11. Cheng, Q., et al., *Removal of group B streptococci colonizing the vagina and oropharynx of mice with a bacteriophage lytic enzyme.* Antimicrob. Agents Chemother., 2005. **49**: 111–117.
12. Loeffler, J.M., D. Nelson, and V.A. Fischetti, *Rapid killing of Streptococcus pneumoniae with a bacteriophage cell wall hydrolase.* Science, 2001. **294**: 2170–2172.
13. Loessner, M.J., et al., *Complete nucleotide sequence, molecular analysis and genome structure of bacteriophage A118 of Listeria monocytogenes: Implications for phage evolution.* Molec. Microbiol., 2000. **35**(2): 324–340.
14. Bateman, A. and N.D. Rawlings, *The CHAP domain: a large family of amidases including GSP amidase and peptidoglycan hydrolases.* Trends Biochem. Sci., 2003. **28**: 234–237.

Chapter 19

General M13 Phage Display: M13 Phage Display in Identification and Characterization of Protein–Protein Interactions

Kirsten Hertveldt, Tim Beliën, and Guido Volckaert

Abstract

In M13 phage display, proteins and peptides are exposed on one of the surface proteins of filamentous phage particles and become accessible to affinity enrichment against a bait of interest. We describe the construction of fragmented whole genome and gene fragment phage display libraries and interaction selection by panning. This strategy allows the identification and characterization of interacting proteins on a genomic scale by screening the fragmented "proteome" against protein baits. Gene fragment libraries allow a more in depth characterization of the protein–protein interaction site by identification of the protein region involved in the interaction.

Key words: Phage display, panning, affinity selection, fragmented whole genome library, gene fragment display library, protein–protein interaction, functional annotation, protein–protein interaction characterization.

1 Introduction

Identification and characterization of protein interactions are often important tools in the unraveling of protein functions. Besides yeast two-hybrid *(1)*, M13 phage display *(2)* and interaction selection by affinity purification *(3)* have shown to be powerful tools in the investigation of molecular interactions. Both techniques can be performed on an individual partner, as well as on a whole genome scale where they allow identification of interacting proteins by sequencing the corresponding DNA regions. In phage display and interaction selection by affinity purification (panning), proteins or polypeptides of interest (in a manner

Martha R. J. Clokie, Andrew M. Kropinski (eds.), *Bacteriophages: Methods and Protocols, Volume 2: Molecular and Applied Aspects, vol. 502*, © 2009 Humana Press, a part of Springer Science+Business Media
DOI 10.1007/978-1-60327-565-1_19 Springerprotocols.com

analogous to the yeast two-hybrid system, we call it "the prey") are exposed as fusion proteins to one of the surface proteins (ordinarily gp*3*, gp*8* or gp*6*) of filamentous phage particles. In the phage particle, there is a physical link between the genotype (the DNA inside the phage particle) and the phenotype (the protein exposed on the phage coat). Subsequently, pools of phages exposing different proteins/peptide sequences are subjected to successive rounds of interaction selection against a target molecule of interest (analogously to yeast two-hybrid, we call it "the bait") (**Fig. 19.1**). Each round of interaction selection comprises

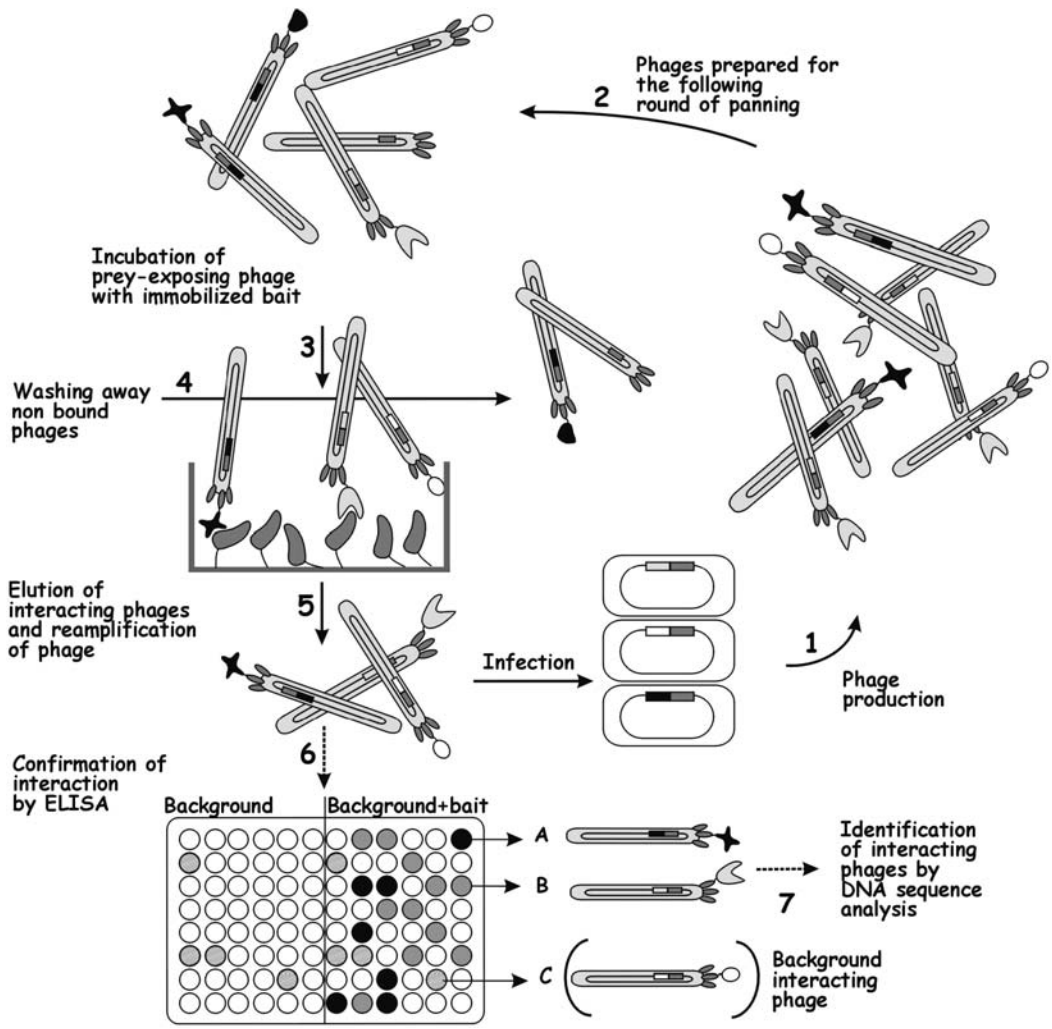


Fig. 19.1. Panning strategy. Phage production from a specific library in *E. coli* (1) yields a phage pool (2) which is incubated with an immobilized bait (3). Non-binding phages are removed by washing (4) and interacting phages are eluted and re-amplified in *E. coli* (5). Generally, after three to five rounds of panning, phages are analyzed for bait interaction by phage ELISA (6). Bait-interacting phages (A and B) are identified by DNA sequence analysis. C represents a phage interacting with the background.

incubation of the prey-exposing phage pool with the bait, removal of non-binding phage, elution, and amplification of bound phages in F^+ *Escherichia coli*. The strength of the panning procedure results from the powerful combination of interaction selection and biological amplification. A large pool of prey-exposing phages (up to 10^{13} phages/mL) can be screened for their ability to interact with the bait. Interacting phages may be very rare in the original phage pool, but can be recovered and enriched by successive rounds of affinity selection and amplification.

Besides applications in immunology (isolation of monoclonal antibodies) *(4)* (see also **Chapter 20**) and epitope identification *(5)*, phage display and selection by panning are helpful in protein annotation by identification and characterization of protein–protein interactions. Peptide libraries, cDNA libraries, and genomic fragment libraries have shown promising results in genome-wide protein interaction screening *(6, 7, 8)*. Gene fragment libraries allow more detailed characterization of the interaction by delineation of the amino acid region involved in bait binding *(9)*. Screening of large libraries of protein variants allows identification of crucial amino acids in the bait–prey interaction, unraveling the influence of amino acids in altered binding characteristics *(10)* and in protein stability *(11)*.

Interaction selection can be performed against a wide variety of molecules (protein, DNA, RNA, carbohydrate, lipids). Target proteinaceous bait molecules may be purified natural proteins which are directly attached to a matrix support or recombinant tagged proteins which can be immobilized by means of protein tags. The latter approach may lead to a better presentation, since conditions of passive absorption can cause protein denaturation *(12)*. Bait molecules can be immobilized on Petri dishes, on microtiter plates, immunotubes, on magnetic beads coated with streptavidin, Ni-NTA or agarose beads in an affinity matrix.

In the protocols presented below, we focus on the construction of fragmented whole genome DNA and fragmented gene libraries in phagemid pHOS31 *(13)*, derived from pHEN1 *(14)*, which allows low valency prey exposition at the N-terminal of gp*3*. The protocols can be used in combination with other gp*3* display phagemids as well. For selection of low-affinity binders, the use of phagemids with gp*8*-fused exposition for polyvalent display (e.g., phagemid vector pG8H6 *(15)*) is recommended.

The basic panning protocol is adapted from original protocols described (*see (3, 16, 17)*). For more information about the biology of filamentous phages and different panning protocols (*see (18, 19)*). Methods and protocols in antibody phage display are described by O'Brien and Aitken *(20)* (*see* also **Chapter 20**). Here, we focus primarily on a general protocol for identification and characterization of protein–protein interactions. In the panning protocol, we describe direct coating with purified protein.

In the notes, we add small modifications to the basic protocol in case tagged proteins and indirect protein anchorage are used. Anchorage by tags can improve bait presentation in a native conformation. To optimize bait–prey interaction selection and to minimize prey selection against the background, we recommend that the matrix and means of immobilization of the bait be alternated in successive panning rounds. The panning protocol can be applied in combination with available peptide libraries [e.g., Ph.D.-12™ Phage display peptide Library kit (New England Biolabs, Ipswich, MA; http://www.nbsc.com/Main.asp)] to screen for bait-interacting peptides and identification of interacting proteins by motif searches *(6)*. Phage ELISA is performed to discriminate between specific binders and phages selected against the background.

2 Materials

2.1 Construction of Whole Genome/Gene Fragment Display Libraries

T4 DNA polymerase, DNA polymerase I (Klenow fragment), and T4 polynucleotide kinase are from Westburg (Leusden, The Netherlands; http://www.westburg.nl/). T4 DNA ligase is from Promega Corp. (Madison, WI; http://www.promega.com). Calf intestine alkaline phosphatase (CIP) and restriction enzymes are from Roche Applied Sciences (Indianapolis, IN; http://www.roche-applied-science.com/). Commercial kits for plasmid isolation and for purification of DNA fragments are from Qiagen (Valencia, CA; http://www1.qiagen.com/). Square Bio-assay dishes (245 × 245 × 25 mm) and sterile 96-microwell flat-bottomed plates are from Nalge Nunc International (http://www.nalgenunc.com/). *E. coli* XL1 Blue MRF' competent cells are from Stratagene (La Jolla, CA; http://www.stratagene.com/) or prepared as described *(21)*. The disposable nebulizer device is from Lifecare Hospital Supplies Ltd (Walsall, United Kingdom). Microcon YM-30 centrifugal devices are from Millipore Corporation (Billerica, MA; http://www.millipore.com/). Gene Pulser™ and electroporation cuvettes (1 mm gap) are from Bio-Rad (Hercules, CA; http://www.bio-rad.com/).

2.2 Media and Solutions

1. Luria-Bertani medium (LB medium): 10 g/L Bacto-tryptone, 5 g/L Bacto-yeast extract, and 10 g/L NaCl in deionized H_2O. Autoclave.
2. 2x TY medium (2TY medium): Bacto-tryptone to 16 g/L, Bacto-yeast extract to 10 g/L, and NaCl to 5 g/L in deionized H_2O. Autoclave.

3. LB agar and 2TY agar: LB medium or 2TY medium supplemented with 15 g/L agar.
4. SOC medium: Bacto-tryptone at 20 g/L, Bacto-yeast extract at 5 g/L, NaCl at 0.5 g/L, 10 m*M* $MgSO_4$, 10 m*M* $MgCl_2$, and 20 m*M* glucose (the latter three components are added from 100X stock solutions, filter-sterilized, separately).
5. 10XTE buffer: 100 m*M* Tris, 10 m*M* EDTA (pH 7.8) with HCl.
6. EB buffer: 10 m*M* Tris (pH 8.6) with HCl.
7. 1% (w/v) agarose gel: 1 g agarose in 100 mL TAE buffer (40 m*M* Tris–HCl (pH 7.2), 500 mM sodium acetate, and 50 mM EDTA).
8. 20% D (+) glucose (Glc), filter-sterilized.
9. Ampicillin (Ap), 100 mg/mL in Milli-Q water (Milli-Q water = deionized water with a resistivity of 18 MΩ-cm) stock solution, aliquot and store at −20 °C. The working concentration corresponds to 100 μg/mL (Ap^{100}): dilute 1 μL stock solution/1 mL medium.

2.3 Panning of Phage Display Libraries

HRP/anti-M13 monoclonal conjugate is from GE Healthcare (http://www.gehealthcare.com/). MaxiSorb flat-bottomed 96-microwell plates for panning and ELISA and sterile round-bottomed 96-microwell plates for growth of cell cultures are from Nagle Nunc International. A commercial kit for the isolation of plasmid DNA (miniprep scale) is from Qiagen or Biognost Research (Heule, Belgium; http://www.biognost.be/). The 0.22 and 0.45 μm syringe driven filter units (Millex filter units with PVDF or PES membranes) are from Millipore. Ni-NTA HisSorb plates are from Qiagen. Protein of interest is purified protein from a natural source or the soluble fraction of recombinant protein (e.g., E- and His_6-tagged protein produced by cloning in pQE-EN *(22)*). 50 mL Sterile conical centrifuge tubes with caps are from Sarstedt or 50 mL Falcon™ tubes from BD Biosciences (San Jose, CA; http:// www.bdbiosciences.com/).

2.4 Media and Solutions

Luria-Bertani medium (LB medium), 2xTY medium (2TY medium), LB agar and 2TY agar, and 20% Glc, Ap^{100} as mentioned in **Section 2.1**.

1. M9 minimal medium (for 1 L) is prepared by mixing 100 mL 10xM9 salts, 1 mL $MgSO_4$ (1 *M*), 1 mL $CaCl_2$ (0.1 *M*), 1 mL thiamine.HCl (1 *M*), 10 mL 20% Glc, and 887 mL of autoclaved water. To obtain solid medium, supplement with 15 g agar. 10xM9 salts (128 g $Na_2HPO_4.7H_2O$, 30 g KH_2PO_4, 5 g NaCl, and 10 g NH_4Cl for 1 L water) and the $MgSO_4$ solutions were autoclaved separately. The other components are sterilized individually by filtration through

a 0.22 μm filter (Millipore Corporation). Store the thiamine.HCl solution at 4 °C.

2. Kanamycin, 35 μg/mL working concentration (Km^{35}). Prepare a 1000X stock solution (35 mg/mL) in Milli-Q water, freeze aliquots.
3. Tetracycline, 10 μg/mL working concentration (Tc^{10}). Prepare a 1000X stock solution (10 mg/mL) in 50% ethanol, in −20 °C (wrap the stock solutions and media supplemented with Tc^{10} in foil and store in the dark).
4. 2TY-K: 2TY medium supplemented with Km^{35} (dilute stock solution to 1X in the medium).
5. 2TY-AK: 2TY medium supplemented with Ap^{100} and Km^{35} (dilute stock solution to 1X in the medium).
6. 2TY-AG: 2TY medium supplemented with Ap^{100} (dilute stock solution to 1X in the medium) and Glc to 2% (w/v).
7. PEG/NaCl: 20% (w/v) polyethylene glycol (PEG) 6,000 and 2.5 *M* NaCl. Dissolve (brief heating to 65 °C may be necessary) and autoclave. Store at 4 °C.
8. 5XPBS: 750 m*M* NaCl, 40 m*M* Na_2HPO_4, and 7.8 m*M* KH_2PO_4 (pH 7.4–7.6). Autoclave solution and then store at room temperature.
9. PBS: dilute 5XPBS to 1xPBS with Milli-Q water and adjust pH to 7.4–7.6.
10. PBST: 1XPBS with 0.1% Tween-20: 1 mL Tween-20/L PBS. Tween-20 is very viscous, hence pipette it very slowly. Store solution at room temperature.
11. Carbonate coating buffer: 0.1 *M* sodium hydrogen carbonate buffer (pH 9.6): Mix solutions of 0.1 *M* Na_2CO_3 and 0.1 *M* $NaHCO_3$ until pH 9.6. Store at room temperature.
12. Blocking buffer: PBS–3% BSA: bovine serum albumin (BSA, Sigma-Aldrich, St. Louis, MO; http://www.sigmaaldrich.com/) prepared as a 3% (w/v) solution in PBS.
13. Elution buffer: 100 mM triethylamine (TEA): 140 μL triethylamine/10 mL Milli-Q water, prepare freshly. TEA is highly flammable. Work in a hood and read safety instructions. TEA should be treated as hazardous waste. Be sure you know the procedures of chemical waste disposal specified by your institution.
14. Neutralization buffer: 1.0 *M* Tris–HCl (pH 7.4). HCl is irritating and corrosive.
15. ABTS (2,2′-azino-di-(3-ethylbenzthiazoline-6-sulfonate)) Microwell Peroxidase substrate system (Kirkegaard & Perry Laboratories, Gaithersburg, MD; http://www.kpl.com/). Store at 4 °C.

2.5 Strains and Helper Phage

F pilus-expressing *Escherichia coli*, e.g., *E. coli* XL1 Blue MRF′ (XL1) or *E. coli* XL1 TG1 (TG1). Grow cells regularly under

selective conditions for the retention of the F′ episome, required for the production of the bacterial F-pilus which is essential for phage infection. In XL1-Blue cells the F′ episome carries a tetracycline-resistance gene thus growth on LB/Tc^{10} selects for the presence of the F′ plasmid. In TG1 cells, the F′ episome encodes enzymes involved in proline biosynthesis (*proAB*$^+$), which have been deleted in the host chromosome. Only the bacteria that carry the F′ plasmid are proline prototrophs and grow on M9 minimal medium. TG1 cells should be routinely grown on this medium to maintain the F′ plasmid.

Helper phage VCSM13 from Stratagene carries a kanamycin-resistance gene.

3 Methods

3.1 Standard Protocols for the Construction of Whole Genome/Gene Fragment Libraries

3.1.1 Fragmentation of DNA and end Repair of DNA Fragments

1. Purified target DNA (*see* **Note 1**) is randomly sheared by nebulization at 1.5 bar argon pressure in 2 mL ice-cold 1xTE-50% glycerol using a disposable nebulizer device (*see* **Note 2**).
2. Precipitate DNA by adding 0.1 volume of NaOAc (3 *M*, pH 5.2) and 2.5 volumes of cold EtOH (95%). Mix carefully. Recover the DNA by centrifugation at 16,000*g* for 30 min. Decant the supernatant and wash the pellet with 0.5 mL cold 70% ethanol. Spin down again and remove the ethanol. Dissolve the air-dried pellet in 40 µL EB buffer.
3. End-repair of DNA fragments is performed for 30 min at room temperature (22 °C) in T4 DNA polymerase reaction buffer (supplemented with BSA solution provided by the supplier), containing T4 DNA polymerase (0.5–1 U/1 µg DNA), Klenow DNA polymerase (0.5–1 U/1 µg DNA), and 33 µM of each dNTP.
4. Purify DNA by a standard commercial DNA purification kit (*see* **Note 3**) and phosphorylate DNA fragments with T4 polynucleotide kinase (4 U/1 µg DNA) at 37 °C for 1 h in kinase buffer (provided by the manufacturer) supplemented with 1 mM rATP.
5. Purify DNA fragments and elute in 50 µL EB (*see* **Note 4**).

3.1.2 Restriction Digestion and Dephosphorylation of Vector DNA

1. Display vector pHOS31 (10–50 µg) (*see* **Note 5**) is subjected to *Sma*I restriction digestion (2 U/1 µg vector DNA) in 200 µL reaction buffer at 25 °C for 2 h. Purify the DNA (*see* **Section 3.1.1** (4)).
2. Dephosphorylate vector DNA with calf intestinal phosphatase (CIP; 0.2–1 U/µg DNA) at 37 °C for 1 h. Purify the DNA (*see* **Note 6**).

3.1.3 Ligation of DNA Fragments in Phagemid Vector and Transformation to E. coli

3.1.3.1 Small-Scale Trial Ligations and Transformations

1. Determine the DNA concentration of both vector and insert fragments by comparison with appropriate DNA markers (typically λ DNA digested with *Pst*I and small-size fragment marker if the desired DNA fragments are < 700 bp) on a 1% (w/v) agarose gel.
2. Set up a series of test ligation reactions to estimate optimal molar vector:insert ratios (e.g., 2:1, 1:2, 1:5, 1:10, 1:50) with a fixed concentration of vector DNA (50 ng/10 μL) with 10 U T4 DNA ligase/1 μg vector DNA in 100 μL. Incubate the ligation reactions overnight at 16 °C. Heat-inactivate the ligase enzyme at 70 °C for 10 min and precipitate the DNA using NaOAc/EtOH (*see* **Section 3.1.1** (2)). Resuspend pellet in 3 μL Milli-Q water and electroporate 1–3 μL (15–50 ng vector) into 40 μL competent cells (*see* **Note 7**). Alternatively, electroporate 2 μL (~10 ng vector) of the overnight ligation mixture into 40 μL competent cells without further purification of the ligation mixture.
4. Perform electroporations to 40 μL XL1-Blue MRF′ cells in 1 mm cuvettes at 1.7 kV (electroporation conditions should be according to the manufacturer's recommendations). After electroporation, quickly add 1 mL prewarmed SOC medium and transfer to a small culture glass tube. Incubate at 37 °C for 1 h.
5. Plate 10-fold serial dilutions onto 2TY-AG plates to determine the number of transformants. Incubate overnight at 37 °C. Count cells and calculate the number of transformants (*see* **Note 8**).

3.1.3.2 Large-Scale Library Ligations and Transformation to *E. coli*

1. Perform large-scale ligation reactions in line with the planned size of the library and with the number of transformants from the test ligations under optimal vector:insert ratio. Prepare the number of ligation reactions required for the library construction (on average 1–5 μg cut vector DNA in a total volume of 100–200 μL). Incubate the ligation reactions overnight at 16 °C.
2. Heat-inactivate the ligase at 70 °C for 10 min and precipitate DNA by NaOAc/EtOH (*see* **Section 3.1.1** (2)). Resuspend the pellet in Milli-Q water to a concentration of ~50 ng vector DNA/1 μL.
3. Use highly competent cells (competence of 10^9–10^{10} transformants/μg supercoiled DNA) if high amounts of clones are required (*see* **Note 8**). Electroporate the ligated material in fractions of 1–2 μL (corresponding to max 50–100 ng vector DNA; transformation efficiency decreases with electroporations of > 100 ng/40 μL cells) to 40 μL electrocompetent *E. coli* XL1 Blue MRF′ (Stratagene) (*see* **Section 3.1.4**). After electroporation, quickly add 1 mL

prewarmed SOC medium and transfer to a small culture glass tube. Repeat this procedure until all the ligation product has been transformed. 20–100 electroporations may be required to obtain large libraries. Incubate at 37 °C for 1 h.

4. Spin down the cells between 3,300 and 6,000 g for 10 min, remove the supernatant, and resuspend in 4–8 mL 2TY-AG. Plate 10-fold serial dilutions of an aliquot onto small 2TY-AG Petri dishes to determine the size of the library. Plate out the total library on four large square Bio-assay plates. Incubate overnight at 37 °C. Count cells and calculate the number of transformants, i.e., the number of primary clones in the library.
5. Toothpick 96 individual colonies for clone analysis. Subsequently, scrape all cells from the plates after addition of 2TY-AG. Supplement half of the volume with glycerol to 20% end concentration. Divide in 500 µL aliquots and store at −70 °C as primary library stocks. Pellet the other half of the volume for plasmid isolation (by a commercial kit) and store the DNA at −20 °C.
6. Analyze the library by PCR clone analysis to analyze size of fragments and the percentage of clones with insertion (> 80% is desirable) (*see* **Note 9**).

3.2 Standard Protocols for Panning of Phage Display Libraries

1. Grow *E. coli* TG1 (XL1-Blue) in a glass tube containing 4 mL 2TY medium at 37 °C to $OD_{600} = 0.5–0.6$ (*see* **Note 10**).
2. Add helper phage VCSM13 (8×10^9 pfu/ml, MOI 20:1) (*see* **Note 11**) and incubate for 15–30 min at 37 °C for infection (no shaking).

3.2.1 Helper Phage Production

3. Dilute cells in 100 mL 2TY/Km^{35} (in a 1 L Erlenmeyer flask with good aeration) and grow overnight, shaking. Next day, precipitate phages by PEG/NaCl (**Section 3.3.2**).

3.2.2 PEG/NaCl Precipitation of (helper) phage particles

1. Spin down cells at 6,000 g for 10 min at 4 °C. If cells are spun down at 3,300 g for 20 min (which also works well), this step should be performed twice. In the latter case, transfer the supernatant to a fresh tube and re-spin to pellet the remaining cells.
2. Transfer clear supernatant containing the phage particles to a clean tube and add 1/5 volume of the mixture PEG/NaCl. Mix by inversion and incubate on ice for at least 30 min.
3. Collect phages by centrifugation at $10,000\,g$ for 20 min ($3,300\,g$ for 30 min also works well) at 4 °C. Discard the supernatant.
4. Resuspend the phage pellet in PBS (1/100 volume of the supernatant) by pipetting up and down with a filter tip. Collect in Eppendorf tubes and spin down the remaining cells at 16,000 g for 2 min at 4 °C.

5. Filter the phage containing solution through a 0.45 or 0.22 μm filter and store helper phages at −20 °C (or for long-term storage in 15% glycerol at −70 °C). Determine the phage concentration by titration (*see* **Procedure 3.2.3**).

3.2.3 (Helper) Phage Titration

1. Prepare 10-fold dilutions of the phage preparation in PBS. These dilutions can be prepared in, e.g., small glass culture tubes, Eppendorf tubes or sterile 96-microwell plates.
2. Add an equal volume of TG1 ($OD_{600} = 0.5–0.6$) and incubate at 37 °C for 25 min (no shaking). Include a control containing only TG1 and PBS without phage (*see* **Note 12**).
3. Plate cells on 2TY agar and incubate overnight at 37 °C. Count plaques and calculate the corresponding titer from the plate which has between 100 and 1,000 colonies.

3.2.4 Initial Phage Production from Phage Libraries for Panning

1. Inoculate an amount of library TG1/XL1-Blue bacteria (from a glycerol stock) corresponding to at least 10X up to 100X the library size. Inoculate in 50 mL 2TY-AG in a 250 mL Erlenmeyer flask. The inoculum should not exceed 0.1 OD_{600}. If XL1-Blue is grown, medium should be supplemented with Tc^{10} (*see* **Note 13**).
2. Grow cells with shaking (270 rpm) at 37 °C to an $OD_{600} = 0.5–0.6$. This will bring the cells in mid-log growth phase, so that they express the F-pilus.
3. When $OD_{600} = 0.5–0.6$ is reached, transfer 5 mL ($\sim 2 \times 10^9$ bacteria) to a 50 mL centrifuge tube containing 4×10^{10} pfu VCSM13 (MOI = 20 : 1). Incubate at 37 °C for 30 min, standing with occasional gentle agitation.
4. Spin the cells for 10 min at 3,300 g and remove the supernatant. Resuspend the bacterial pellet in 25 mL 2TY-AK and transfer to a 250 mL flask. Grow with shaking (270 rpm) overnight at 30 °C (*see* **Note 14**).
5. Precipitate phage particles with PEG/NaCl (*see* **Procedure 3.3.2**). The standard yield is about $2–10 \times 10^{12}$ phages from a 25 mL culture. Check the phage titer (*see* **Procedure 3.2.3**, but plate cells on 2TY-AG (instead of 2TY) agar after infection) (*see* **Note 15**).

3.2.5 Selection of Bait-Specific Phage Preys: Panning and Amplification of Phage

1. MaxiSorb flat-bottomed 96-microwell plates are coated overnight at 4 °C (or for 2 h at room temperature) with an appropriate concentration of purified protein in coating buffer (*see* **Note 16**). Alternatively, tagged proteins can be immobilized by specific anti-tag-antibodies or other means of anchorage (*see* **Notes 17** and **18**). The number of wells required depends on the diversity of the library. Ideally, the phage concentration should not exceed 10^{13} phage/ml and

the total number of phage should exceed the library diversity by 1,000-fold *(10)*. Thus, for a diversity of 10^{10}, 10^{13} phage should be used and, using a concentration of 10^{13} phage/ml and 100 µL/well, 10 wells are required. At the same time, reserve an equal number of uncoated wells (only coating buffer) as a negative control. Seal the wells. Upscaling from 0.1 mL volumes to 1 mL can be performed in MaxiSorb immunotubes (Nunc).

2. Discard the bait solution and wash the wells three times with PBST and three times with PBS. Fill to the brim, let stand for 30 s to 1 min and tap out the buffer.
3. Block the wells with blocking buffer (PBS–3% BSA) for 2 h at room temperature (or overnight at 4 °C). Fill to the brim (*see* **Note 19**).
4. Preblock library phages in PBS–3% BSA
5. Remove the blocking buffer and wash three times with PBST and three times with PBS.
6. Add 200 µL of library phage ($\sim 10^{10}$–10^{11} input phages) in PBS/3% BSA (*see* **Note 20**) to each of the coated and uncoated wells. Seal the wells. Incubate for 2 h at room temperature with gentle shaking on an orbital plate shaker (e.g., on a Heidolph unimax 1010 shaker (Heidolph Laboratory Equipment, Schwabach, Germany)).
7. Remove the phage solution and wash five times with PBST and five times with PBS (*see* **Note 21**).
8. Elute the phage (output phage of both bait coated wells and control wells) with 200 µL elution buffer for 10 min at room temperature (covered with parafilm or a lid) on an orbital shaker. Transfer the 200 µL solution to a glass tube/50 mL tube filled with 100 µL of neutralization buffer (*see* **Note 22**).
9. A 5–10 µL of the output phage solution is used for phage titration (*see* **Procedure 3.2.3**, but plate cells on 2TY-AG (instead of 2TY) agar after infection; *see* **Note 23**). The eluate must be diluted at least 10-fold for toxicity reasons of TAE.
10. To the bulk of the eluted phage solution (290–295 µL) of bait-coated wells only 4 mL of actively growing TG1 or XL1-Blue cells ($OD_{600} = 0.5$–0.6) are added. Incubate for 30 min at 37 °C (no shaking). The eluate must be diluted at least 10-fold for toxicity reasons of TEA.
11. Transfer the 4 mL (phagemid infected culture) to 36 mL 2TY-AG (containing 10^{10} pfu VCSM13/mL). Incubate with slowly shaking (~200 rpm/min) for 2 h at 37 °C. Spin the culture at 3,300 g for 10 min, discard the supernatant, and resuspend the pellet in 40 mL 2TY-AK. Incubate overnight at 30 °C (*see* **Note 14**) with shaking (*see* **Note 24**).
12. Isolate the phage particles by PEG/NaCl precipitation. Check the titer of library phages by phage titration (*see*

Procedure 3.2.3). Repeat the sorting cycles until the enrichment ratio has reached a maximum (*see* **Note 23**). Use phage immediately or store filtered solutions (0.2 μm) at 4 °C for maximum 1 week. Enrichment is generally observed in round 3 or 4 and performing six selection rounds is seldom necessary.

3.2.6 Analysis of Selected Prey Molecules: Production of Phagemids in MTP

1. Inoculate colonies from the plates used to titer output phage in 100 μL 2TY-AG in round-bottomed 96-well plates and grow with shaking overnight at 37 °C (*see* **Note 25**).
2. Use a 96-well sterile transfer device to inoculate 5 μL/well from this plate to a round-bottomed 96-well plate containing 150 μL 2TY-AG per well. Transfer a proportional inoculum to a 4 mL tube to follow the OD_{600}. Grow to an OD_{600} = 0.5–0.6 at 37 °C, shaking. To the wells of the master plate, add 50 μL of 60% glycerol per well and store at −70 °C.
3. To each well add 50 μL 2TY-AG containing 2×10^9 pfu VCSM13 phage (MOI = ~20 : 1). Incubate for 30 min at 37 °C (no shaking).
4. Spin the plate at 500 g for 10 min and remove the supernatants using a multichannel pipet.
5. Resuspend the pellets in 150 μL 2TY-AK. Grow the cells overnight at 30 °C.
6. Spin the plate at 500 g for 10 min and use 50 μL supernatant per well for phage ELISA (*see* **procedure 3.2.7**, step 5).

3.2.7 Analysis of Selected Prey Molecules: Phage ELISA

1. Coat MaxiSorb microtiter plates with purified bait protein in 100 μL coating buffer/well. Cover with the lid and incubate at room temperature for 2 h (or overnight at 4 °C). Prepare a control plate with coating buffer only (*see* **Note 26**).
2. Wash the plates three times with PBST and three times with PBS.
3. Block (fill to the brim) bait and control plates with blocking solution for 2 h at room temperature.
4. Wash three times with PBST and three times with PBS.
5. Add 50 μL of phage supernatant (*see* **Procedure 3.2.6**, step 6) to a total volume of 100 μL PBS–3% BSA to both plates. Incubate at room temperature for 2 h.
6. Wash four times with PBST and four times with PBS.
7. Add 100 μL of HRP/anti-M13 monoclonal conjugate (1:5,000 in PBS–3% BSA). Incubate at room temperature for 1 h.
8. Wash four times with PBST and four times with PBS.
9. Add 100 μL of the ABTS Microwell Peroxidase substrate and monitor colour development for 45 min at 405 nm in a microplate reader. Depending on the signal, discriminate between preys selected against the bait and against the background. Identify bait-specific clones by DNA sequence analysis

of the corresponding plasmids (by standard plasmid isolation after inoculation from corresponding wells from the master plate and DNA sequence analysis).

4 Notes

1. For whole genome display libraries, target DNA is purified genomic DNA; for gene fragment libraries, target DNA is a specific PCR product. Due to loss of DNA during fragmentation, verification and subsequent purification steps, we recommend to start with enough DNA (~25–50 μg genomic DNA for bacterial genomes and at least 10–20 μg of PCR product).
2. The pursued fragment size depends on the desired length of exposed fragments. Both gp*3* and gp*8* phagemid systems can be used. The multivalency of the gp*8* phagemid system can enhance selection of low affinity interactions. The gp*8* phagemid system, however, is more prone to deletion than gp*3* phagemid systems when large proteins are displayed. This may be the result of size-exclusion effects during phage assembly *(23)*. For genome wide interaction screening, fragments between 100 and 1,000 bp are often used. Alternatively, two separate library pools are prepared, a small-fragment (~100–500 bp), and a longer fragment library (~400–800 bp). Screening of gene fragment libraries is generally performed to select small interaction regions and critical residues, hence small fragment sizes are desired.

   Methods for random DNA fragmentation for cloning in phage(mid) vectors are similar to protocols described *(24)* to construct DNA sequencing libraries. Popular methods for random fragmentation are hydrodynamic shearing methods, such as sonication *(25)* and nebulization *(26,27)*. Enzymatic cleavage methods include fragmentation by partial digestion with, e.g., *Cvi*J1 under "relaxed" conditions or controlled degradation with DNase I *(28)*.

   We fragment DNA by nebulization (http://www.genome.ou.edu/protocol_book/protocol_partII.html; *(29)*) using a disposable nebulizer device. Different types of nebulizers are in use, e.g., BioNeb® Cell disruption system (Glas-Col Apparatus Co., Terre Haute, IN; http://www.glascol.com) and Nebulizers from Invitrogen (Carlsbad, CA). Cheaper alternatives are nebulizer devices from medical supply stores. These should be adapted to minimize dispersion (aerosol leakage) of the liquid. Maximum allowed pressure should be checked. If you intend to shear DNA that could possibly contain pathogenic agents (e.g., pathogenic viral DNA), you should take

precautions to ensure that the aerosolized solution is not inhaled. Work in an appropriate safety cabinet. Conditions for nebulization are determined rather empirically and are ordinarily tested on small samples, prior to preparative nebulization. Nebulization time and pressure influences the size of the DNA fragments obtained *(24)*. Depending on the size of the target DNA and the desired fragment size, we vary the nebulization time (the medical nebulizing device we use has a maximum allowed pressure of 1.5 bar), but variation of pressure is also recommended *(27)*. To generate a quite even distribution of DNA fragment sizes, nebulization is performed on ice or in an ice-water bath. Alternatively, the nebulization device is incubated with intervals of 2 min after each 2 min of nebulization. If during long nebulization times buffer volume reduces too much (especially for fragmentation of small DNA fragments), add buffer to the original volume. Follow the extent of fragmentation during the nebulization process by taking ~500 ng samples and analysis on a 1% TAE agarose gel. We obtain fragments of 400–600 bp after 4 min nebulization of genomic DNA (~6 Mb) at 1.5 bar. The genomic DNA is isolated on Nucleobond® AXG columns (Macherey-Nagel, Düren, Germany; https://www.macherey-nagel.com/) with Nucleobond® buffer set III. Smaller target DNA (e.g., PCR products for gene fragment libraries) requires higher pressure or longer nebulization times (up to 20–30 min *(30)*). Depending on the size distribution of the sheared fragments (which in general can be rather narrow) and personal preference one may decide to further purify DNA fragments of a specific size range by preparative gel electrophoresis or other techniques for size fractionation (e.g., preparative electrophoresis through a 10 cm 2.5% agarose column on a model 230A HPEC system from Applied Biosystems).

DNase I fragmentation in the construction of gene fragment display libraries is described *(9)*. Sonication as a means of random fragmentation in the construction of display libraries has been described *(31)* and *(32)*, for the construction of whole genome and gene fragment libraries, respectively.

3. We use commercial DNA purification kits for removal of enzymes and purification of DNA and will refer in further steps to this procedure as "Purify DNA." We obtain good results with Qiagen PCR purification and Qiagen nucleotide removal kits. Prior to DNA purification with commercial kits, the enzyme is inactivated as described by the supplier. Alternatively, purify by standard phenol/chloroform extraction followed by NaOAc/EtOH precipitation (see *(24)*).

4. Check DNA yield and fragment sizes on TAE agarose gel. In general, after nebulization and subsequent enzymatic reactions and purification steps, the DNA recovery is about 25–50%. If a narrow fragment size range is desired, DNA fragments can be extracted from agarose gel (using commercial kits).
5. Conditions for *Sma*I digestion and cloning in display vector pHOS31 *(13)* are described here. Display vector pHOS31 allows display of proteins fused to gp*3*. For cloning purposes, we isolate plasmid DNA by means of QIAGEN Tip-100 from *E. coli* XL1 Blue MRF' transformed with pHOS31. Take into account vector purification yields.
6. If after examination on agarose gel, vector digestion is not complete, gel extraction of the linearized vector is necessary.
7. Include the following controls: vector with ligase (to check for efficiency of the dephosphorylation reaction), vector without ligase (to check for the presence of uncut vector), and (optionally) vector with a control blunt-end restriction fragment (to check efficiency of ligation).
8. Library size is a major determinant in successful selection. Test ligation reactions allow to estimate library sizes obtained with different vector:insert ratios and they provide useful information regarding the scale of ligation required to achieve a desired library complexity. The optimal ratio should be applied for large-scale ligation. Calculate the number of colony-forming units/μg DNA. The required number of clones of your library of interest can be calculated according to the formula of Clarke and Carbon *(33)*. However, take into account that the calculated number of clones should be multiplied by 18, because only $1/3 \times 1/3 \times 1/2$ of the clones will have an insert in the correct orientation and in frame with both the signal sequence and the gene 3 sequence. Realize that DNA fragments generated by nebulization and treated for end-repair are much less efficiently cloned than genuine restriction fragments. High insert to vector ratios may be required.
9. To perform clone analysis PCR, toothpick 96 clones and transfer to a sterile microtiter plate containing 100 μL 2TY-AG. Incubate for 2 h at 37 °C. Samples of 2.5 μL cell culture from the microtiter plate are transferred into 20 μL PCR mixture for PCR analysis with 0.15 U SuperTaq. The mixture further contains 0.6 μ*M* of each primer (M13forward and M13reverse), 20 mM Tris–HCl (pH 8.3), 50 mM KCl, 0.1% Tween 20, 2 mM $MgCl_2$, and 0.2 m*M* of each dNTP. The PCR program consists of the following 25 cycles of DNA amplification: DNA denaturation at 95 °C for 30 s (1 min in the first cycle), primer annealing at 55 °C for 30 s, and polymerization (72 °C, 1 min/kb).

10. Cells grown at 37 °C to $OD_{600} = 0.5–0.6$ express F-pili. These cells can be stored at 4 °C (on ice) for several hours without significant loss of pili.
11. $OD_{600} = 0.5$ corresponds to approximately 4×10^8 cells/mL. A multiplicity of infection (MOI) of 20 phages: one bacterium corresponds to 8×10^9 phages/mL.
12. If dilutions are prepared in a sterile cell growth plate (microtiter plate), add an additional amount of cells to the petri dish while spreading cells to allow good cover of the whole surface of the petri dish (~400 µL total volume).
13. When using frozen aliquots of the library, you should first thaw an aliquot and check the number of viable cells by plating dilutions.
14. Lowering temperature to 30 °C during phage production can improve protein expression and exposition. Efficient infection only takes place at 37 °C.
15. Phage can be stored at 4 °C without much loss of titer. The exposed proteins and polypeptides, however, may be degraded proteolytically by contaminating proteases. For selection, it is recommended that phages are used immediately. Solutions that are not used immediately should be filter-sterilized (0.22–0.45 µm filter) to prevent bacterial growth.
16. Bait proteins for coating are usually at 1–10 µg/100 µL. Some baits require 100 µg/100 µL. Protein amounts and coating conditions (coating buffer, temperature) should be tested individually by an appropriate ELISA. We usually coat the plates with 1 µg bait/100 µL in carbonate buffer (PBS can be a good alternative). The concentration of immobilized bait can be varied in the different panning rounds. Immobilization of higher amounts of bait in the first panning round(s) may facilitate selection of library phage which are present at low concentration. Decreasing the amount of bait coated will allow selection for high affinity binders.
17. Direct coating may influence protein conformation and hamper proper interaction with phage exposed proteins. Therefore, indirect coating conditions by means of antibodies against tagged proteins or Ni-NTA against His_6-tagged proteins are used as indirect means of immobilization of proteins of interest. For tagged recombinant proteins, coat wells of a MaxiSorb plate with anti-tag antibody (1 µg/100 µL coating buffer). Coat for 2 h at room temperature (or overnight at 4 °C). Wash three times with PBST and three times with PBS. Block with 3% (w/v) BSA in PBS. At the same time block an equal number of uncoated wells as a negative control. Wash three times with PBST and three times with PBS. Incubate wells with an appropriate amount of purified recombinant protein (~1 µg) or coat directly from a cleared cell lysate

(22). Coating can be performed overnight at 4 °C or for 2 h at room temperature. Proceed as indicated in protocol 3.2.5.1, step 4.

18. For coating of Ni-NTA HisSorb plates with His_6-tagged proteins from cleared bacterial lysate, the recommended ELISA conditions are used (Qiagen). Bait protein is captured from the lysate in Ni-NTA-plates (also incubate negative control lysates in wells). Wash four times with PBST and four times with PBS. Add 200 µL library phages preblocked with PBS–3% BSA to each of the bait and control wells and incubate at room temperature with gentle shaking. Washing and elution is performed as described in Procedure 3.2.5, steps 7–8. Be aware that anchorage by means of the Ni-NTA may result in selection of library phages against Ni-NTA. This will depend on the presence of His-rich sequences in the display library.
19. An alternative blocking agent is skimmed milk powder (2% (w/v) in PBS *(17)*). One can also purchase blocking reagents such as Superblock from Pierce Chemical Company (Rockford, IL; http://www.piercenet.com/).
20. Some bait–prey interactions require specific buffer conditions (e.g., pH, salts). In these situations, specific buffers should be used for interaction selection.
21. The number and length of washing steps and the presence of detergents may be varied depending on the stringency preferred. Detergents such as Tween-20 (0.1–0.5% (v/v)) are often added to reduce non-specific binding of the phage. In the first round, washes can be limited to six in total and increased in subsequent rounds.
22. Triethylamine (TEA) is flammable and irritating. Wear gloves and work in a hood. Prepare the dilution freshly to have a high pH of the solution. Elution times longer than 10 min can damage the phage. Alternatively, elution can be performed under acidic conditions with 50 mM glycine–HCl (pH 2) followed by neutralization with one volume of 200 mM sodium phosphate buffer (pH 7.5) *(34)*.
23. Prepare serial dilutions of eluted phage, infect TG1, and spread the cells on 2TY-AG plates to determine the amount of phage eluted. Calculate the % bound phage per round = the number of output phage divided by the number of input phage in each round. The enrichment in round n is calculated as % bound phage in round n divided by the % bound phage in round $n-1$. In a successful panning, each round enriches for phages that bind the bait, and the pool is eventually dominated by specific binders. Additionally, determine the enrichment ratio over the background = the number of phage eluted from a well coated with the bait divided by the number of phage eluted from a control well. This enrichment ratio over background gives an indication of specific selection

against the bait. The enrichment ratio over the background usually peaks at a particular round, from which clones should be analyzed *(10)*.

24. Alternatively, infected cells (4 mL culture) can be spread on large round (14.5 cm Ø) 2TY-AG agar plates to allow cell growth on plate to reduce bias resulting from growth competition between clones in liquid medium. After overnight incubation at 30 °C, scrape of colonies and produce phages as in procedure 3.2.4.
25. Fix the sealed plate in the incubator. Place it as far as possible from the ventilator to avoid evaporation.
26. Indirect immobilization of the bait protein of interest can be performed by anchorage by means of Ni-NTA (if the protein contains a His_6 tag) or by specific antibodies against a protein tag. Perform coating and blocking steps as explained in notes 17 and 18. Continue this procedure from step 5.

Acknowledgements

K. Hertveldt is a postdoctoral scientist of the Flemish FWO (Fonds voor Wetenschappelijk Onderzoek-Vlaanderen). T. Beliën holds a predoctoral fellowship of the IWT (Instituut voor de aanmoediging van Innovatie door Wetenschap en Technologie in Vlaanderen). The author's projects are financially supported by the Research Council of the K.U. Leuven (grants OT/98/20, OT/04/30, and OT/05/45), the Flemish FWO (research grants G.017297N, G.0114.01, and G.0308.05), the GBOU (Generic Basic Research at the Universities) grant "Xylafun" (IWT-010081), and the EU contract BIO4-CT97-2294B4.

References

1. Fields, S. and Song, O. K. (1989) A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245–246.
2. Smith, G. P. (1985) Filamentous fusion phage: Novel expression vectors that display cloned antigens on the surface of virion. *Science* **228**, 1315–1317.
3. Parmley, S. F. and Smith, G. P. (1988) Antibody-selectable filamentous fd phage vectors: Affinity purification of target genes. *Gene* **73**, 305–318.
4. Hoogenboom, H. R., de Bruine, A. P., Hufton, S. E., Hoet, R. M., Arends, J. W. and Roovers, R. C. (1998) Antibody phage display technology and its applications. *Immunotechnol.* **4**, 1–20.
5. Scott, J. K. and Smith, G. P. (1990). Searching for peptide ligands with an epitope library. *Science* **249**, 386–390.
6. Kay, B. K., Kasanov, J. Knight, S. and Kurakin, A. (2000) Convergent evolution with combinatorial peptides. *FEBS Lett.* **480**, 55–62.
7. Crameri, R. and Kodzius, R. (2001) The powerful combination of phage surface display of cDNA libraries and high throughput screening. *Comb. Chem. High. Throughput Screen.* **4**, 145–55.
8. Tong, A. H., Drees, B., Nardelli, G., Bader, G. D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., Quondam, M., Zucconi, A., Hogue, C. W., Fields, S., Boone, C. and Cesareni, G. (2002)

A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* **295**, 321–4.
9. Wang, L. F., Du Plessis, D. H., White, J. R., Hyatt, A. D. and Eaton, B.T. (1995) Use of a gene-targeted phage display random epitope library to map an antigenic determinant on the bluetongue virus outer capsid protein VP5. *J. Immunol. Methods* **178**, 1–12.
10. Sidhu, S. S., Lowman, H. B., Cunningham, B. C. and Wells, A. (2000) Phage display for selection of novel binding peptides. *Methods Enzymol.* **328**, 333–363.
11. Forrer, P., Jung, S. and Plückthun, A. (1999) Beyond binding: Using phage display to select for structure, folding and enzymatic activity in proteins. *Curr. Opin. Struct. Biol.* **9**, 514–520.
12. Butler, J., Ni, L., Nessler, R., Joshi, K. S., Suter, M. Rosenberg, B., Chang, J., Brown, W. R. and Cantarero, L. A. (1992) The physical and functional behaviour of capture antibodies adsorbed on polystyrene. *J. Immunol. Methods* **150**, 77–90.
13. Hertveldt, K., Robben, J. and Volckaert, G. (2002) In vivo selectively infective phage as a tool to detect protein interactions: evaluation of a novel vector system with yeast Ste7p-Fus3p interacting proteins. *Yeast* **19**, 499–508.
14. Hoogenboom, H. R., Griffiths, A. D., Johnson, K. S., Chiswell, D. J.,Hudson, P. and Winter, G. (1991) Multi-subunit proteins on the surface of filamentous phage: methodologies for displaying antibody (Fab) heavy and light chains. *Nucleic Acids Res.* **19**, 4133–4137.
15. Jacobsson, K. and Frykberg, L. (1996) Phage display shot-gun cloning of ligand-binding domains of prokaryotic receptors approaches 100% correct clones. *Biotechniques* **20**, 1070–1081.
16. McCafferty, J., Griffiths, A. D., Winter, G. and Chiswell, D. J. (1990) Phage antibodies: filamentous phage displaying antibody variable domains. *Nature* **348**, 552–554.
17. Marks, J. D., Hoogenboom, H. R., Bonnert, T. P., McCafferty, J., Griffiths, A. D. and Winter, G. (1991) *J. Mol. Biol.* **222**, 581–597.
18. Kay, B. K., Winter, J. McCafferty, J. (ed.) (1996) *Phage display of peptides and proteins.* Academic Press, San Diego, CA.
19. Barbas III, C. F., Burton, D. R., Scott, J. K. and Silverman, G. J. (ed.) (2001). Phage display. A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
20. O'Brien, P. M. and Aitken, R. (ed.) (2002) *Antibody phage display. Methods and protocols.* Humana, Totowa, NJ.
21. Tung, W. L. and Chow, K. C. (1995). A modified medium for efficient electrotransformation of *E. coli*. *Trends Genet.* **11**, 128–129.
22. Hertveldt, K., Dechassa, M. L., Robben, J. and Volckaert, G. (2003) Identification of Gal80p-interacting proteins by *Saccharomyces cerevisiae* whole genome phage display. *Gene* **27**, 141–9.
23. Marciano, D. K., Russel, M. and Simon, S.M. (1999) An aqueous channel for filamentous phage export. *Science* **284**, 1516–9.
24. Sambrook, J. and Russell, D. W. (ed.) (2001) *Molecular Cloning, a Laboratory Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
25. Deininger, P. L. (1983) Approaches to rapid DNA sequence analysis. *Anal. Biochem.* **135**, 247–263.
26. Bodenteich, A., Chissoe, S., Wang, Y.-F. And Roe, B. A. (1994) Shotgun cloning as the strategy of choice to generate templates for high-throughput dideoxynucleotide sequencing, in *Automated DNA sequencing and analysis* (Adams, M. D., Fields, C. and Venter, J., ed.), Academic Press, San Diego, CA, pp. 42–50.
27. Hengen, P. N. (1997) Shearing DNA for genomic library construction. *Trends Biochem. Sci.* **22**, 273–274.
28. Anderson, S. (1981) Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res.* **9**, 3015–3027.
29. Okpodu, C. M., Robertson, D., Boss, W. F., Togasaki, R. K. and Surzycki, S. J. (1994) *Biotechniques* **16**, 154–159.
30. Rudgers, G. W. and Palzkill, T. (2001) Protein minimization by random fragmentation and selection. *Protein Engineering* **14**, 487–492.
31. Jacobsson, K. and Frykberg, L. (1995) Cloning of ligand-binding domains of bacterial receptors by phage display. *Biotechniques* **18**, 878–885.
32. Di Niro, R., Ferrare, F., Not, T., Bradbury, A. R., Chirdo, F., Marzari, R. and Sblattero, D. (2005) Characterizing monoclonal antibody epitopes by filtered gene fragment phage display. *Biochem. J.* **388**, 889–894.
33. Clarke, L. and Carbon, J. (1976) A colony bank containing synthetic Col El hybrid plasmids representative of the entire *E. coli* genome. *Cell* **9**, 91–9.
34. Kay, B.K., Kasanov, J., Yamabhai, M. (2001) Screening phage-displayed combinatorial peptide libraries. *Methods* **24**, 240–6.

Chapter 20

Isolation of Monoclonal Antibody Fragments from Phage Display Libraries

Mehdi Arbabi-Ghahroudi*, Jamshid Tanha* and Roger MacKenzie

Abstract

Techniques developed over the past 20 years for the display of foreign peptides and proteins on the surfaces of filamentous bacteriophages have been a major driving force in the rapid development of recombinant antibody technology in recent years. With phage display of antibodies as one of its key components, recombinant antibody technology has led to the development of an increasing number of therapeutic monoclonal antibodies. Antibody gene libraries are fused to a gene encoding a phage coat protein. Recombinant phage expressing the resulting antibody libraries in fusion with the coat protein are propagated in *Escherichia coli*. Phage displaying monoclonal antibodies with specificities for target antigens are isolated from the libraries by a process called panning. The genes encoding the desired antibodies selected from the libraries are packaged within the phage particles, linking genotype and phenotype. Here, we describe the application of this technology to the construction of a phage-displayed single-domain antibody (sdAb) library based on the heavy chain antibody repertoire of a llama, the panning of the library against a peptide antigen and the expression, purification, and characterization of sdAbs isolated by panning.

Key words: Phage display, phage library panning, single-domain antibodies, recombinant antibodies, antibody engineering.

1 Introduction

Display of foreign peptides on the surface of filamentous phages (M13, F1, or fd) as fusions to either the P3 or the P8 coat protein was introduced by George Smith and co-workers *(1–3)* and subsequently applied to larger protein molecules, in particular,

*The two authors contributed equally.

Martha R. J. Clokie, Andrew M. Kropinski (eds.), *Bacteriophages: Methods and Protocols, Volume 2: Molecular and Applied Aspects, vol. 502*, © 2009 Humana Press, a part of Springer Science+Business Media
DOI 10.1007/978-1-60327-565-1_20 Springerprotocols.com

to antibody fragments such as Fabs, scFvs, and sdAbs *(4–16)*. This technology combines characteristics such as genotypic diversity, genotype/phenotype coupling, selection pressure, and clonal amplification *(17)*. Therefore, it is a high throughput screening strategy by which libraries with up to 10^{11} variants can be screened and from which binders with a frequency as low as 1 in 10^6 can be isolated and characterized *(18, 19)*. As a result, it is possible to obtain peptides from random peptide libraries *(20–22)* and antibody fragments from immune antibody libraries *(9,23)*, naïve antibody libraries *(16,24–28)*, and synthetic antibody libraries *(29–31)* with high affinities and specificities for virtually any target antigen. The peptides and antibodies thereby obtained can be affinity-maturated or fine-tuned by various mutagenesis techniques and screening by phage panning against the targets *(32,33)*.

The key elements which determine successful isolation of binders from immune or random libraries are the nature of immune response, library size, and heterogeneity and the choice of phage vectors or phagemid vectors, in combination with different helper phages, which results in various levels of antibody or peptide display. Moreover, the display level is highly dependent on the size and composition of the antibody fragments and their intrinsic folding properties because of the limited potential of the *Escherichia coli* host in terms of the assembly of eukaryotic proteins *(17,34–36)*.

Antibodies or peptides are selected from phage display libraries by a process referred to as biopanning. Typically, pure antigen is adsorbed to solid supports such as ELISA plates *(14, 37, 38)*. As alternative strategies, in situ panning on target antigens on fixed prokaryotic cells *(39)*, fixed mammalian cells *(40)*, living cells *(41,42)*, and phage internalization via cell surface receptors *(43,44)* have been performed. The existence of antigen in its natural state and the possibility of identifying novel binders to unidentified cell surface markers are some of the advantages of the in situ approaches. Following exposure of phage particles to a target, non-specific binders are removed in a washing step and phages bound to target are recovered by elution. The eluted phages are re-amplified in *E. coli* and used in subsequent rounds of panning. Factors such as a gradual reduction in the amount of coated antigen in later rounds of panning, the stringency of washing (long washes should favor the recovery of higher affinity/slow k_{off} binders) and elution conditions, incubation time of phage with target, and the use of a competitive amount of free antigen (below the desired dissociation constant, K_d) will determine the affinity of peptides or antibodies selected by panning. It should be pointed out that affinity is not the only criterion regulating the selection of specific clones and that other factors including availability of epitopes on the target, loss of clones due to

sensitivity to bacterial proteases, and toxicity of some clones to the bacterial host can hinder the selection of the best binders in the library. An important guiding factor for the progress of the selection process is the enrichment factor which is based on the phage yield after panning. Input and output phage numbers are enumerated as colony-forming units (cfu) or plaque-forming units (pfu). The yield is calculated as the number of output phage/number of input phage at each round and compared to the previous round. The enrichment factors vary for each panning experiment but typically are between 10^3- and 10^6-fold after four rounds of panning *(5, 15, 17, 37, 45, 46)*. Other complementary approaches to monitor enrichment for antigen-specific binders are colony-PCR (to check for the presence of antibody genes) and polyclonal and/or monoclonal phage ELISA (to follow the increase in the number of antigen-binding clones) after each round *(47)*.

Screening of the polyclonal mixture of antigen-specific clones is accomplished by a phage-ELISA/cell assay in which phages are prepared from randomly selected individual colonies after the last round of panning and tested for binding to antigen coated on microtiter plate wells or whole cells. Positive clones are characterized at the nucleotide level, by sequencing the peptide or antibody encoding regions in the corresponding phagemid or phage vector. Expression of antibody fragments in soluble form is performed using either a non-suppressor strain of bacteria or subcloning the antibody genes into an appropriate vector. Affinity tag chromatography provides pure protein for determination of binding constants, specificities, aggregation states, and stabilities *(17, 19, 32, 34, 35, 47–49)*.

In this chapter, we describe protocols for the construction of a phage-displayed antibody library, panning of the library, screening of clones following panning, and the expression and characterization of the antibodies isolated from the library. Specifically, we describe the construction of an sdAb library from the lymphocytes of a llama immunized with a peptide antigen, the isolation from the library of sdAbs specific for the peptide, and characterization of the antibodies. The sdAbs are derived from a unique type of antibody produced by camels and llamas, namely, heavy chain antibodies comprised only of heavy chains *(50)*.

2 Materials

2.1 Cells, Phages, and Vectors

1. TG1 electroporation-competent cells (Stratagene, La Jolla, CA; http://www.stratagene.com/).
2. M13KO7 helper phage (New England Biolabs, Ipswich, MA; http://www.neb.com/).
3. pSJF2 expression vector *(51)*.
4. pJT1 phagemid vector (**Fig. 20.1**).

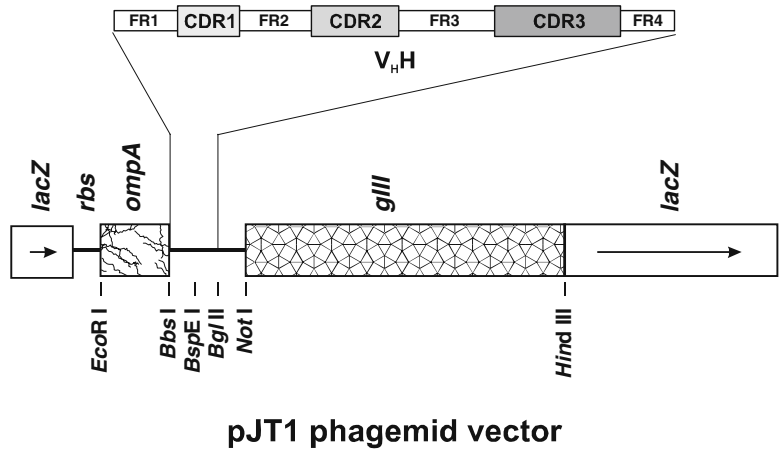


Fig. 20.1. Schematic representation of the pJT1 phagemid vector used to construct the V_HH phage display library. The vector was constructed from the expression vector pSJF2 *(51)* by inserting bacteriophage Fd *gIII* between *Bgl*II and *Hin*dIII restriction sites employing standard cloning techniques. The V_HH library was cloned between *Bbs*I and *Bgl*II sites. This places the V_HH genes precisely between the OmpA signal peptide gene (*ompA*) and *gIII*. The arrow in the diagram shows the direction of *lacZ* expression in the parent vector. *rbs*, ribosome binding site; CDR, complementarity-determining region; FR, framework region.

2.2 Primers

Primers were purchased from Sigma Genosys (The Woodlands, TX; http://www.sigmaaldrich.com/Brands/Sigma_Genosys.html).

1. CH2FORTA4:5′-CGCCATCAAGGTACCAGTTGA-3′ *(4)*.
2. CH2B3-F: 5′-GGGGTACCTGTCATCCACGGACCAGCTGA-3′.
3. VHBACKA6: 5′-GATGTGCAGCTGCAGGCGTCTGGRGGAGG-3′ *(4)*.
4. CVHHP35BACK: 5′-CAGGCTCAGGTACAGCTGGTGGAGTCTGG-3′.
5. VHBACKA6Bbs: 5′-TATGAAGACACCAAGCCGATGTGCAGCTGCAGGCGTCT-3′.
6. CVHHP35Bbs-R: 5′-TATGAAGACACCAAGCCCAGGCTCAGGTACAGCTGGTG-3′.
7. VHBgl: 5′-TATAGATCTTGAGGAGACGGTGACCTG-3′.
8. VHBbs: 5′-TATGAAGACACCAGGCCGATGTGCAGCTGCAGGCG-3′.
9. VHBbs1: 5′-TATGAAGACACCAGGCCCAGGCTCAGGTACAGCTGGTG-3′.
10. VHBam: 5′-TATGGATCCTGAGGAGACGGTGACCTG-3′.
11. −96gIII: 5′-CCCTCATAGTTAGCGTAACGATCT-3′.
12. M13RP: 5′-CAGGAAACAGCTATGAC-3′.
13. M13FP: 5′-GTAAAACGACGGCCAGT-3′.

2.3 Media and Solutions

1. SOC, LB (Luria-Bertani), 2xYT and induction media (Terrific Broth with no salts) *(52)*.
2. Wash solution: 10 mM Tris buffer (pH 8.0) and 154 mM NaCl.
3. Sucrose solution: 10 mM Tris buffer (pH 8.0), 1 mM EDTA, and 25% sucrose.
4. Shock solution: 10 mM Tris buffer (pH 8.0) and 0.5 mM $MgCl_2$.
5. Starting buffer: 10 mM HEPES (*N*-[2-hydroxyethyl] piperazine-*N*′-[2-ethanesulfonic acid]) buffer, 10 mM imidazole, and 500 mM NaCl (pH 7.0).
6. Elution buffer: 10 mM HEPES buffer, 500 mM imidazole, and 500 mM NaCl (pH 7.0).
7. Neutralizing buffer: 1 M Tris buffer (pH 7.4).
8. Sterile PBS (phosphate-buffered saline) (*52*).
9. MPBS: 2% (w/v) skim milk in PBS.
10. 0.05% and 0.1% PBST: 0.05% and 0.1% (v/v) Tween 20 in PBS.
11. NaPi buffer: 6.7 mM Na_2HPO_4, 3.3 mM $NaH_2PO_4.H_2O$, 150 mM NaCl, and 0.5 mM EDTA (pH 7.0).
12. HBS-E buffer: 10 mM HEPES buffer (pH 7.4), 150 mM NaCl, and 3 mM EDTA. This buffer can be purchased from GE Healthcare, Baie d′urfé, QC, Canada; http://www.gehealthcare.com. (Piscataway, NJ). If not purchased from GE Healthcare, it should be thoroughly degassed before use.
13. Sterile PEG/NaCl solution: 20% (w/v) polyethylene glycol 6,000 or 8,000, 2.5 M NaCl.
14. 100 mM triethylamine: 35 μL of 7.18 M triethylamine in 2.5 mL ddH_2O, made fresh daily.
15. Sterile double distilled water (ddH_2O).

Solutions were sterilized by autoclaving (*52*).

2.4 Reagents and Consumables

1. Ampicillin.
2. Kanamycin.
3. Triethanolamine.
4. IPTG (isopropyl-β-D-thio-galactopyranoside).
5. Surfactant P20 (GE Healthcare).
6. Amine coupling kit containing *N*-hydroxysccinimide (NHS), *N*-ethyl-*N*′-(3-dimethylaminopropyl)carbodiimide hydrochloride (EDC) and ethanolamine (GE Healthcare)).
7. TMB peroxidase substrate and H_2O_2 (Kirkegaard & Perry Laboratories, Inc. (KPL), Gaithersburg, MD; http://www.kpl.com/).
8. Biotinylated FMDV22 peptide Biotin-GGGGKYGENAVTN VRGDLQVLAQKAA-RTLPTSF (Sheldon Biotechnology Centre, Montreal, QC, Canada; http://www.mcgill.ca/sheldon/).

9. Streptavidin.
10. Restriction enzymes and T4 DNA ligase (New England Biolabs, Ipswich, MA; http://www.neb.com/); Taq DNA polymerase.
11. Durex™ 13 × 100 mm borosilicate glass tubes (VWR Scientific).
12. HRP-conjugated anti-M13 monoclonal antibody (GE Healthcare).
13. QIAamp RNA Blood Mini™ kit (Qiagen, Inc., Valencia, CA; http://www1.qiagen.com/).
14. First-Strand cDNA Synthesis™ kit (GE Healthcare).
15. QIAquick Gel Extraction™ kit (Qiagen, Inc.).
16. QIAquick PCR Purification™ kit (Qiagen, Inc.).
17. LigaFast™ Rapid DNA Ligation System (Promega Corp., Madison, WI; http://www.promega.com).
18. A 0.2 μm GP Express™ Plus Membrane filtration system (Millipore Corp., Billerica, MA; http://www.millipore.com/).
19. A 5 mL HiTrap™ Chelating HP column (GE Healthcare)
20. Superdex 75 gel filtration column (GE Healthcare).
21. 96 Microtiter wells and Maxisorp™ strip wells (VWR Scientific).
22. Electroporation and disposable cuvettes.
23. CM5 sensor chips (GE Healthcare)).

2.5 Equipment

1. Agarose gel electrophoresis equipment.
2. SDS-PAGE (sodium dodecyl sulfate-polyacrylamide gel electrophoresis) equipment.
3. DNA sequencing equipment.
4. ÄKTA FPLC purification system (GE Healthcare).
5. Thermal cycler.
6. Multi-well (ELISA) plate reader.
7. Sorval high speed and swinging bucket bench-top (RT6000B Refrigerated) centrifuges or their equivalents.
8. Microfuge.
9. ND-1000 spectrophotometer (NanoDrop Technologies, Inc., Wilmington, DE; http://www.nanodrop.com/) or a similar instrument.
10. Cell density meter (Fisher Scientific) or any regular spectrophotometer.
11. BIACORE 3000 (GE Healthcare)) or other Biacore instrument with similar capabilities.
12. MicroPulser™ electroporator (Bio-Rad Laboratories, Hercules, CA; http://www.bio-rad.com) or a similar one.
13. Incubators for bacterial growth on plates and liquid media.

3 Methods

3.1 Library Construction

A sdAb library was constructed from the lymphocytes of a llama immunized with a peptide, CKYGENAVTNV-RGDLQVLAQKAARTLPTSF, derived from a capsid protein of the Foot and Mouth Disease Virus; the N-terminal cysteine was added for peptide synthesis purposes. The library was constructed according to the protocol shown in **Fig. 20.2** and detailed below.

1. Isolate total lymphocyte RNA from 2 mL llama blood using QIAamp RNA Blood Mini^{TM} kit according to the manufacturer's instructions. Measure RNA concentration and purity at ODU_{260} and ODU_{280}, respectively *(52)*.
2. Use a total of 3–5 μg RNA in 20 μL water to synthesize cDNA in a total volume of 33 μL, using First-Strand cDNA

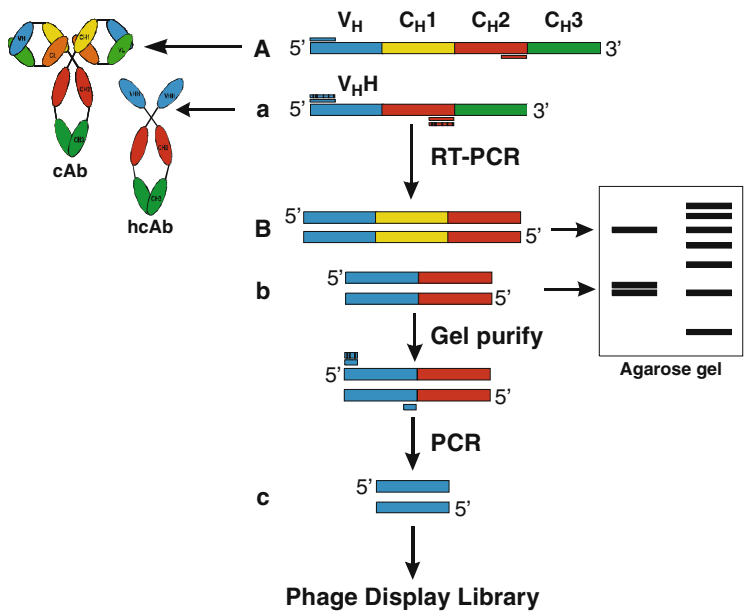


Fig. 20.2. Schematic representation of the steps involved in isolating a repertoire of V_HH genes. cDNA was synthesized by reverse transcription (RT) from mRNA of conventional (A) and heavy chain (a) antibodies using two different C_H2-specific primers. The cDNA was subsequently converted to dsDNA using the C_H2-specific primers and two FR1-specific primers. Three PCR products, one derived from "A" (B) and two smaller products derived from "a" (b), were obtained and resolved on a 2% agarose gel. The two shorter fragments were gel-purified and subjected to a second round of PCR to amplify the V_HH gene repertoires using two FR1- and a FR4-specific primers. The primers introduced appropriate flanking restriction endonuclease sites for subsequent cloning of the V_HH repertoire as a phage display library. For simplicity only coding sequences, showing the variable (V) and constant (C_H) regions, of the mRNAs are shown. cAb, conventional antibody; hcAb, heavy-chain antibody.

Synthesis™ kit and C_H2-specific primers, CH2FORTA4 and CH2B3-F, according to the manufacturer's instructions (*see* **Note 1**).

3. Perform test PCRs using various amount of the cDNA reaction mix ranging in volume from 1 to 5 μL using the primer mix CH2FORTA4, CH2B3-F, and framework 1 (FR1)-specific primers VHBACKA6 and CVHHP35BACK:

dNTPs (2.5 mM each)	4 μL
10x PCR buffer	5 μL
CH2FORTA4 and CH2B3-F (10 pmol/μL each)	0.5 μL
VHBACKA6 and CVHHP35BACK (10 pmol/μL each)	0.5 μL
cDNA mixture	1–5 μL
Taq DNA polymerase (5 units/μL)	0.5 μL
H_2O	34.5–38.5 μL
Total volume	50 μL

   Place the reaction tubes in a thermal cycler and synthesize dsDNA from cDNA on a program consisting of a preheating step at 94 °C for 5 min followed by 30 cycles of 94 °C for 30 s, 55 °C for 30 s, and 72 °C for 1 min.
4. Analyze 5 μL of the PCR products on a 2% agarose gel *(52)*. Identify the cDNA volume that gives the best yield in terms of amplifying V_HH genes (**Fig. 20.2**) and perform the PCR experiment for the remaining cDNA mixture under the same conditions (*see* **Note 2**). Gel-purify the V_HH bands on a 2% agarose gel using the QIAquick Gel Extraction™ kit. Pool the DNAs and calculate the concentration of DNA based on ODU_{260} measurements *(52)*.
5. Use 10–20 ng of the amplified cDNA/reaction tube in a second PCR to amplify V_HH genes as described above using the FR1-specific primers, VHBACKA6Bbs and CVHHP35Bbs-R, and the FR4-specific primer VHBgl (*see* **Note 3**). Carry out a total of 20 PCRs. Purify the amplified products with QIAquick PCR Purification™ kit.
6. Digest about 20 μg of V_HH DNA with *Bbs*I overnight. Analyze a small sample on a 1% agarose gel to ensure that it is the proper size. Purify the digested product with two QIAquick PCR Purification™ columns, elute each in 100 μL sterile ddH_2O, and measure the DNA concentration (the yield of the purified material should be around 80%). Re-digest the sample with *Bgl*II overnight, analyze on an agarose gel, purify and measure the concentration as above (*see* **Note 4**).
7. Digest about 60 μg phagemid vector pJT1 with *Bbs*I overnight and examine on a 1% agarose gel to ensure that

the vector is completely linearized. Gel-purify the linearized product using several QIAquick Gel Extraction™ columns and elute in 100 µL ddH_2O for each column (use a maximum of 400 mg of gel per spin column). Measure the DNA concentration (the yield should be about 30–40%). Re-digest the sample with *Bgl*II overnight, analyze on a 1% agarose gel, purify with QIAquick PCR Purification™ columns, and measure the concentration (*see* **Note 5**).

8. Perform ligation in a total volume of 100 µL with a vector to insert molar ratio of 1:2 using LigaFast™ Rapid DNA Ligation System:

Digested vector	20 µg
Digested V_HH insert	3.5 µg
T4 DNA ligase buffer (2x)	50 µL
T4 DNA ligase (3 units/µL)	8 µL
Sterile ddH_2O	Adjust to 100 µL

   Incubate at room temperature for 60 min (*see* **Note 6**).
9. Purify the ligated materials using a QIAquick PCR Purification™ kit on a single column and elute the DNA in a final volume of 35 µL sterile ddH_2O. Use a few microliters to measure the DNA concentration.
10. Transform 50 µL of electrocompetent TG1 cells with 3 µL of the purified ligated material as described *(53)* using a MicroPulser™ electroporator or an equivalent instrument. Transfer the electroporated cells into a tube containing 1 mL SOC medium and incubate for 1 h at 37 °C and 180 rpm. Repeat the transformation for the remaining DNA (i.e., a total of 10 transformations).
11. Pool the transformed cells, take a small aliquot, and carry out 10^3-, 10^4- and 10^5-fold dilutions in 2xYT. Spread 100 µL of the diluted cells on 2xYT agar medium containing 100 µg/mL ampicillin (2xYT/Amp) and incubate overnight at 32 °C. In the morning, use the plates to determine the functional size of the library as described in **Section 3.2**.
12. Amplify the library by transferring the transformed cells into 500 mL of 2xYT/Amp/2% glucose and incubating overnight at 220 rpm and 37 °C.
13. In the morning, centrifuge the cells at 5,000 *g* for 20 min at 4 °C. Discard the supernatant and resuspend the cells in 50 mL 2xYT/Amp/2% glucose. Make dilutions of the cells in 2xYT, measure the absorbance at ODU_{600}, and use this value to calculate the cell density (number of cells/mL) in the stock solution (1 $ODU_{600} = 10^9$ cells). Add 50 mL 70% glycerol to the cell stock, make several aliquots of 10^{10} bacterial cells/vial and freeze the cells at −80 °C (*see* **Note 7**).

3.2 Determining the Functional Size of the V_HH Library

1. Count the colonies on the titer plates (**Section 3.1, step 11**) and determine the total library size.
2. Carry out colony PCR on the colonies from the titer plates in a total volume of 15 μL. Prepare a master mix for 50 PCRs:

10x PCR buffer	80 μL
dNTPs (2.5 mM each)	64 μL
−96gIII (10 pmol/μL)	16 μL
M13RP (10 pmol/μL)	16 μL
Taq DNA polymerase (5 units/μL)	8 μL
H_2O	616 μL

   Aliquot 15 μL volumes from the master mix in 50 PCR tubes. Touch single colonies from the titer plates with sterile toothpicks (or a P10 pipette tip) and swirl in the PCR tubes. Place the reaction tubes in a thermal cycler and perform PCR with a program consisting of a preheating step at 94 °C for 5 min followed by 30 cycles of 94 °C for 30 s, 55 °C for 30 s, and 72 °C for 1 min and a final step of 72° for 7 min.
3. Apply a few microliters of the PCR mix on 1% agarose gels to identify the clones with full insert (∼ 600 bp) (*see* **Note 8**). Purify the remaining PCR mix for the clones with full insert with a QIAquick PCR Purification™ kit and determine the DNA concentration. Sequence the clones and identify those expressing V_HH sequences *(50, 54)* (*see* **Note 9**). Determine the functional library size by multiplying the percentage of the 50 clones with V_HH sequences by the total library size.

3.3 Production of Phage Library

1. Thaw on ice 10^{10} bacterial cells from the amplified library (**Section 3.1, step 13**) and inoculate 300 mL 2xYT/Amp/1% glucose. Grow to an ODU_{600} of 0.4–0.5 by incubating the cells at 37 °C and 220 rpm.
2. To infect, add M13KO7 helper phage to the culture at a phage to cell ratio of 20:1 (10^{12} pfu), incubate at 37 °C for 5 min without shaking and 0.5–1 h with shaking.
3. Centrifuge the infected bacteria at 5,000 *g* for 10 min at 4 °C and gently resuspend the pellet in 30 mL of 2xYT/Amp/Kan (50 μg/mL kanamycin). Add 270 mL of 2xYT/Amp/Kan and grow overnight at 37 °C and 250 rpm.
4. Next day, pellet the cells at 5,000 g at 4 °C for 20 min and pass the culture supernatant through a 0.2 μm GP Express™ Plus Membrane filtration system.
5. Add 1/5 volume of PEG/NaCl solution, mix well, and incubate in an ice bath for 1 h.
6. Centrifuge the solution at 10,000 *g* for 15 min at 4 °C and resuspend the phage pellet in 2 mL sterile PBS.

7. Centrifuge the phage solution in a microfuge at a maximum speed for 30 s to remove any residual bacterial cell debris. Keep the phage on ice.
8. Prepare exponentially growing TG1 cells by inoculating a 2–3 mL of LB medium in a sterile 15 mL falcon tube with a single colony from a stock plate of TG1 cells (*see* **Note 10**). Incubate at 37 °C in a rotary bacterial shaker at 220 rpm. Remove aliquots from the culture flask at different time intervals and measure the ODU_{600} in a spectrophotometer in disposable cuvettes using LB as the blank. Stop the incubation at $ODU_{600} = 0.4–0.5$ (2–3 h).
9. To determine the titer of the phage, make 10^6, 10^8, 10^{10}, and 10^{12} serial dilutions of phage in PBS, mix 10 μL of each dilution with 100 μL of the exponential-phase TG1 cells. Incubate the cells at room temperature for 15 min and subsequently plate them on 2xYT/Amp medium. In the morning, count the colonies and determine the titer. Phage titers are typically $1–5 \times 10^{13}$ cfu/mL.
10. Store the purified phage at 4 °C for short-term storage, i.e., a few weeks, and in PBS/15% glycerol at −80 °C for long-term storage.

3.4 Panning

In this step, a subtractive panning in which one microtiter well is coated with the subtracting antigen, streptavidin, and another with the target antigen, biotinylated FMDV22 peptide captured by streptavidin, is performed. The phage library is adsorbed on the subtraction well and the supernatant is transferred to the target well-coated with streptavidin–biotinylated peptide complex. The pre-adsorption step in the subtraction well removes streptavidin binders, allowing predominant selection of peptide binders over streptavidin binders in the target well. Panning was performed according to the protocol shown in **Fig. 20.3** and detailed below.

1. Add 100 μL of 100 μg/mL streptavidin in PBS to two Maxisorp™ wells. Seal the wells with parafilm and incubate overnight at 4 °C to coat the wells with streptavidin. Designate one well as the subtraction well and the other as the antigen well.
2. In the morning, discard the streptavidin, blot the wells on a paper towel, and block the wells with 300 μL freshly-made MPBS at 37 °C for 2 h, wells sealed with parafilm (*see* **Note 11**).
3. Discard the blocking solution. To the subtraction well, add 100 μL 10^{12} pfu phage in 2% MPBS and to the antigen well 100 μL of 5 μg/mL biotinylated peptide in PBS. Seal and incubate both wells at room temperature for 1.5 h (*see* **Notes 12 and 13**).

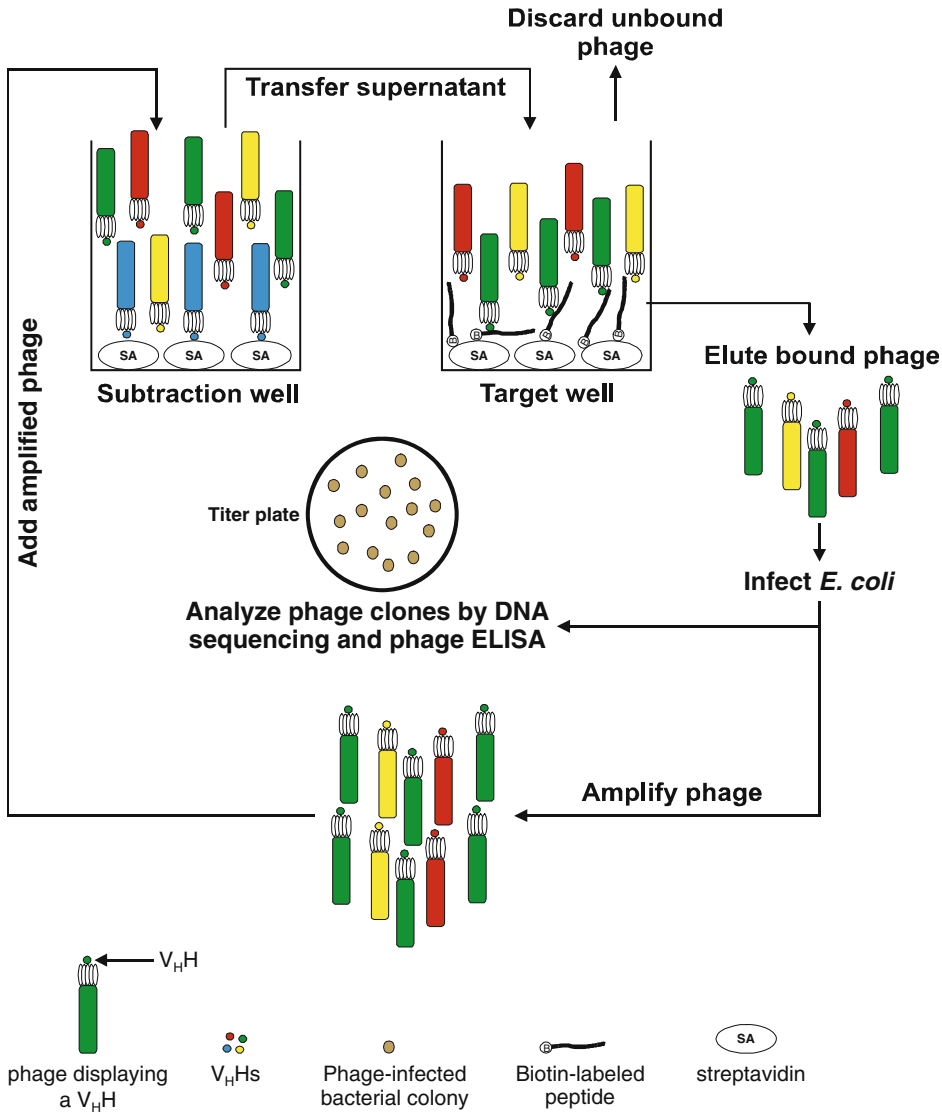


Fig. 20.3. A subtractive panning scheme for isolating peptide-specific V_HHs. The V_HH-displaying phage were first applied to a streptavidin-coated well (subtraction well) to "subtract" streptavidin-specific V_HHs; non-bound phages were then transferred to the target well in which biotin-labeled peptide was captured by streptavidin. Unbound phages were discarded and bound phages were eluted. The eluted phages were amplified by first infecting *E. coli* and then growing the infected cells overnight. The amplified phages were purified and used to initiate another round of panning. The titer of eluted phage was determined for each round, by plating serial dilutions of the infected cells prior to amplification. At the end of the panning, identification of binders was initiated by DNA sequencing and phage ELISA of clones from the titer plates.

4. Empty the antigen well and wash it three times with PBS. Transfer the phage from the subtraction well to the antigen well. Seal and incubate at room temperature for 1.5 h (*see* **Note 14**). Initiate step 8.

5. Discard the unbound phage. Rinse the wells by filling the wells to the brim with 0.1% PBST and discard the wash solution. Repeat the wash step with 0.1% PBST four more times. Wash 10 more times with PBS. After the last wash, blot the well on a paper towel to remove any remaining liquid.
6. Elute the bound phage by adding 100 μL of 100 mM triethylamine. Pipette the content of the well up and down several times and incubate at room temperature for 10 min (*see* **Note 15**).
7. Pipette the content of the well up and down several times; transfer the eluted phage to a cryovial containing 50 μL of 1 M Tris–HCl (pH 7.4) and vortex to neutralize the triethylamine. Keep the tube on ice (*see* **Note 16**).
8. Prepare 10 mL exponentially growing TG1 cells in a sterile 50 mL Falcon tube (*see* **Section 3.3, step 8**). Keep 100 μL of the exponentially growing TG1 cells for the negative control titer in step 9 and infect the remaining with 150 μL of the eluted phage by incubating the mixture of the two at 37 °C for 15 min without shaking and for 1 h with shaking at 220 rpm (any remaining phage may be stored at −80 °C for future reference).
9. Make serial dilutions ($10^{-2} - 10^{-6}$) of the infected cells in 2xYT in 500 μL volumes. Spread 100 μL of each dilution on 2xYT/Amp plates. Also plate 100 μL of the uninfected cells as a negative control. Incubate at 32 °C overnight. Keep the plates parafilm-sealed and stored at 4 °C for clonal analysis (see below and **Section 3.5, step 1**).
10. To the remaining infected cells ($\sim$ 10 mL), add 10^{11} pfu M13KO7 helper phage and incubate for 15 min without shaking and for 1 h with shaking at 220 rpm. Subsequently, add kanamycin at a final concentration of 50 μg/mL and incubate overnight at 37 °C and 250 rpm.
11. In the morning, purify the phage and determine the phage titer (**Section 3.3, steps 4–9**). Use the purified phage to start a new round of panning.

The next rounds of panning are identical to the first except that for each round the input phage is the amplified phage from the previous round and used at a total titer of 10^{11} pfu per well (in Step 3). Typically, three to four rounds of panning are performed to enrich for the binders, although occasionally more rounds may be necessary to obtain enrichment (*see* **Note 17**). The progress of panning can be monitored by colony PCR (**Section 3.2**), sequencing, restriction digest profiling *(47)*, and/or polyclonal phage ELISA *(47)*. In colony PCR, an increase in the number of clones with full-insert over those with no or truncated V_HH is typically observed as the panning progresses. With sequencing and restriction digest profiling, repetition of a few sequences or restriction digest patterns indicate enrichment. In the instance of

a polyclonal phage ELISA, an increase in ELISA signal as a function of panning round is indicative of enrichment for binders. Alternatively, a monoclonal phage ELISA can be performed to monitor enrichment, where an increase in the number of ELISA-positive clones from one round to the next indicates enrichment (**Section 3.5**).

3.5 Monoclonal Phage ELISA

Screening of binders is carried out by monoclonal phage ELISA which is typically performed on the clones from the later rounds of panning. If binder diversity is important, especially when the pool of binders identified by phage ELISA is subjected to a second round of screening for a particular function, phage ELISA can be performed on earlier rounds. The following ELISA format is well suited to the screening of relatively low numbers of clones, less than 25, for example. Cells are grown in 15 mL tubes and the relatively larger amount of phage supernatant obtained in this way allows for multiple assays. For screening larger numbers, a microtiter approach where colonies are grown in microtiter wells in smaller volumes is more feasible *(47)*.

1. Pick single colonies from the titer plates for the eluted phage (**Section 3.4, step 9**) and inoculate 2 mL of 2*x*YT/Amp/0.1% glucose medium in sterile 15 mL disposable Falcon tubes. Grow the cells to an ODU_{600} of 0.5 by incubating them at 37 °C and 220 rpm. Number the colonies on the titer plate for further reference (*see* **Step 10**).
2. Add 10^9 pfu of M13KO7 helper phage to the cells and incubate at 37 °C for 15 min without shaking. Incubate for another 30 min at 37 °C and 250 rpm. Add kanamycin (50 μg/mL) and incubate overnight at 37 °C and 250 rpm.
3. Coat microtiter wells with 100 μL of 5 μg/mL streptavidin in PBS at 4 °C overnight. Coat twice as many wells as the number of the clones being screened.
4. Discard the streptavidin from the microtiter wells and block the wells as described in **Section 3.4, step 2** (*see* **Note 11**).
5. Discard the blocking solution. To half of the wells add 100 μL of 5 μg/mL biotinylated peptide in PBS and to the other half add PBS only. Seal and incubate both wells at room temperature for 30 min.
6. Spin down the cells from step 2 in a bench-top centrifuge at 3,500 rpm for 20 min at 4 °C. Decant the supernatant which contains phage and keep on ice. Add 100 μL of each phage supernatant in duplicates to streptavidin and streptavidin-biotinylated peptide wells. Incubate at room temperature for 1.5 h. Store the remaining phage at 4 °C (*see* **Note 18**).
7. Remove the supernatants, and wash the wells six times with 0.05% PBST. Add 100 μL of a 1:1,000 dilution of HRP-conjugated anti-M13 monoclonal antibody in MPBS to each well and incubate at room temperature for 1 h.

8. Wash the wells as described in **step 7** and blot the plates on paper to remove any remaining liquid. Detect the binding of phage to the cells colorimetrically by adding 100 μL of the TMB peroxidase substrate and H_2O_2 mixture at room temperature for 5–10 min. A blue color should appear.
9. Terminate the reaction by adding 100 μL of 1 M H_3PO_4 (the color will change from blue to yellow) and read the optical density at 450 nm using an ELISA plate reader. (Peptide-specific clones should only show yellow color in streptavidin–peptide wells. The appearance of significant yellow color in streptavidin wells is indicative of the clones being specific for streptavidin or the peptide–streptavidin complex.)
10. Carry out colony PCR (**Section 3.2**) on ELISA-positive clones using the respective parent clones on the reference plate (*see* **Step 1**). Identify the unique V_HH sequences by DNA sequencing and proceed with their cloning and expression as described below.

After five rounds of panning, 48 clones were screened for binding to biotin-FMDV22 peptide–streptavidin complex and streptavidin by phage ELISA. All the clones showed strong binding to the peptide–streptavidin complex but no binding to streptavidin. Forty-one positive clones were subjected to DNA sequencing. Thirty-three had the same V_HH sequence (FMDV22-1) and the remaining eight had a second V_HH sequence (FMDV22-2). The two V_HHs were sub-cloned, expressed, and analyzed for binding (see below).

3.6 Cloning, Expression, Extraction and Purification

3.6.1 Cloning

All the cloning steps were performed essentially as described elsewhere *(52)*.

1. Amplify the V_HH genes from the phagemid vector in a total volume of 50 μL by colony PCR using either VHBbs or VHBbs2 and VHBam primers (*see* **Note 19**). The primers introduce *Bbs*I and *Bam*HI sites at the ends of the amplified fragments.
2. Purify the V_HH genes with a QIAquick PCR Purification™ kit in a final volume of 50 μL water.
3. Cut the purified DNA with *Bbs*I restriction endonuclease and gel-purify with a QIAquick Gel Extraction™ kit in a final volume of 50 μL water. Redigest with *Bam*HI restriction endonuclease and purify with a QIAquick PCR Purification™ kit in 50 μL water.
4. Ligate the cut fragment into *Bbs*I/*Bam*HI-treated pSJF2 expression vector. At the protein level, this results in the addition of C-terminal c-Myc and His_5 tags.

5. Prepare electrocompetent *E. coli* strain TG1 cells *(53)* and use a few microliters of the ligated product to transform the cells as described in **Section 3.1, step 10**. Alternatively, cells can be transformed by chemical transformation *(52)*.
6. Following transformation spread 100 μL of cells on LB/Amp plates and leave the plates with lids half open for 5–10 min on a clean bench. Cover, invert, and incubate overnight at 32 °C.
7. In the morning, perform colony PCR (**Section 3.2**) using M13RP and M13FP primers. Determine the size of the amplified product on a 1% agarose gel. The positive clones (i.e., clones with V_HH genes) should give a size around 650 bp.
8. Confirm the positive clones by further sequencing their V_HH genes as described in **Section 3.2**, using M13RP and M13FP as primers.

3.6.2 Protein Expression and Extraction

V_HH genes are cloned in fusion with the OmpA leader sequence, expressed, and exported to the periplasm. The following extraction protocol based on an osmotic shock method *(55)* is designed to increase the permeability of the outer membrane and release V_HHs from the periplasm, without lysing the cells. Since the endogenous protein content of the periplasm is far less than that of the cytoplasm, the periplasmic extraction step results in partial V_HH purification. It is recommended to keep the fractions from various steps of the extraction at 4 °C, until it is verified by Western blotting which fractions contain the V_HH.

1. Use a single positive clone to inoculate 25 mL of LB/Amp. Incubate in a rotary shaker at 240 rpm overnight at 37 °C.
2. Transfer the entire overnight culture to 1 L of M9 medium supplemented with 5 μg/mL vitamin B1, 0.4% casamino acids, and 100 μg/mL ampicillin. Incubate the culture at 180 rpm for 30 h at room temperature, subsequently supplement with 100 mL of 10x induction medium and 100 μL of 1 M IPTG and incubate for another 60 h.
3. Retain a small aliquot for Western blotting (*see* **Step 7**) and centrifuge the remaining culture at 5,000 g for 20 min at 4 °C in a high-speed centrifuge. Keep the supernatant fraction at 4 °C.
4. Resuspend the pellet in 150 mL wash solution. Centrifuge at 14,000 g for 10 min at 4 °C. Keep the supernatant fraction at 4 °C.
5. Resuspend the pellet in 50 mL sucrose solution and incubate at room temperature for 10 min. Centrifuge at 14,000 *g* for 45 min at 4 °C. Keep the supernatant fraction at 4 °C.
6. Resuspend the pellet in 50 mL ice-cold shock solution and incubate in an ice bath for 10 min. Centrifuge at 14,000 *g* for 25 min at 4 °C. Keep the supernatant fraction at 4 °C.

7. Verify expression by detecting the presence of V_HHs in fractions from steps 3 to 6 by Western blotting against the c-Myc tag *(56)* using anti-c-Myc antibody (*see* **Note 20**). Pool the fractions which contain V_HH and dialyze against 6 L of starting buffer overnight at 4 °C using a dialysis membrane of 10 kDa MW cut-off.
8. Proceed with protein purification.

3.6.3 Purification

The presence of the C-terminal His_5 tag in V_HHs allows for one-step protein purification by immobilized metal affinity chromatography (IMAC) using a 5 mL HiTrap™ Chelating HP column (*see* **Note 21**).

1. Charge the column with Ni^{2+} by applying 30 mL of a 5 mg/mL $NiCl_2.6H_2O$ solution and subsequently wash the column with 15 mL deionized water.
2. Perform purification as described *(56)* using starting buffer, and elute bound protein with a 10–500 mM imidazole gradient.
3. Examine the fractions corresponding to the "eluted" peaks on the chromatogram for the presence and purity of the V_HHs by SDS-PAGE *(57)*. Pool the "V_HH fractions" and dialyze extensively against NaPi buffer. Measure ODU_{280} for determination of protein concentration from molar extinction coefficients *(58)*, add sodium azide at a final concentration of 0.02% and store the V_HHs at 4 °C.

3.6.4 Antibody Affinity Measurements

Standard ELISA methods can be employed to determine the binding specificities of the purified sdAbs. The affinities of the sdAb–antigen interactions can also be estimated by ELISA methods. However, if accurate binding affinities and information of the kinetics of binding are desired, this information can be derived from surface plasmon resonance (SPR) analyses performed with a BIACORE instrument.

1. Isolate monomeric sdAbs prior to SPR analysis using Superdex 75 size exclusion column chromatography (column volume = 25 mL). Equilibrate the column with 50 ml of HBS-E buffer at a pump speed of 0.5 mL/min, inject 200 µL of IMAC-purified V_HH, and collect the monomer peak fraction. Determine the protein concentration.
2. Carry out SPR experiments at 25 °C using a BIACORE 3000 instrument with HBS-E containing 0.005% surfactant P20 as the running buffer.
3. Immobilize streptavidin on a CM5 sensor chip with a surface density of 1,000 response units (RUs). Activate CM-dextran surface with a 7 min injection of a mixture of 50 mM NHS and 200 mM EDC at a flow rate of 5 µL/min. Inject 50 µg/mL streptavidin diluted in 10 mM acetate buffer (pH 4.5) for

3 min and block the surface with a 7 min injection of 1 M ethanolamine (pH 8.5).
4. Capture approximately 80 RUs of biotinylated peptide on the immobilized streptavidin surface by injecting 7 μL of 25 nM biotinylated peptide at a flow rate of 5 μL/min.
5. Analyze V_HH interaction with the peptide using a streptavidin surface as a reference. Inject 10 or 20 μL of more than six different concentrations of monomer V_HH over both the streptavidin and streptavidin–antigen surfaces at a flow rate of 40 μL/min.
6. Analyze the data using BIAevaluation software 4.1 (GE Healthcare). Overlay reference and buffer blank subtracted sensorgrams and determine the amounts of binding at steady state. Calculate the K_Ds by steady-state affinity fitting and from Scatchard plots (*see* **Note 22**).

4 Notes

1. It may sometimes be necessary to optimize the amount of input RNA but generally 3–5 μg total RNA per cDNA synthesis reaction results in a good yield of synthesized DNA by RT-PCR.
2. Three bands are obtained following RT-PCR with average sizes of 900 bp, which corresponds to conventional antibodies, and 690 and 620 bp, which correspond to heavy chain antibodies and contain the V_HH genes *(4)* (*see* also **Fig. 20.2**). The aim of optimizing the PCR is to increase the intensity of the V_HH bands relative to the conventional antibody band. However, differential intensities of the two V_HH bands with respect to each other are routinely observed on agarose gels. Ideally, the PCR may be optimized in four separate reactions, each utilizing a unique primer pair, but we have found this to be unnecessary for immune libraries.
3. Normally, this PCR will result in a single band of about 400 bp. However, if other bands appear, the PCR conditions should be optimized for the concentration of input DNA and $MgCl_2$.
4. Following *Bbs*I/*Bgl*II digestion, we do not observe any significant smaller fragments generated by internal digestion of V_HHs. However, a restriction site analysis of V_HH sequences has shown that over 5% of the V_HH sequences have internal *Bbs*I site.
5. Transformations should be performed with a 1 μg of cut and self-ligated vector. If the number of colonies is too high (there should be less than 10^5 colonies), the vector should be re-cut to reduce the amount of uncut or single cut vector.

6. With a ligation of this magnitude, library sizes of around 10^8 should be obtained. When larger size libraries are required, as in the case for synthetic and naïve libraries, it is advisable to first identify the ligation conditions which give the biggest library size. This can be done by performing small-scale ligations with different total input DNA and molar ratios of insert to vector. Moreover, the scale of the ligation (**Step 8**) and the number of transformations (**Step 10**) need to be significantly increased. In the case of naïve libraries, the amount of input blood (**Step 1**), taken from several individuals, should also be increased as it is the number of antibody-displaying B cells that determines the actual library diversity.
7. For non-immune libraries (naïve, semi-synthetic or synthetic), one should also store the library as purified vector.
8. Generally, with 50 colony PCRs, we observe that over 80% of the clones have V_HH genes, with unique sequences.
9. V_HHs can be distinguished from contaminating V_Hs by the nature of the amino acids at positions 37, 44, 45, and 47 (Kabat numbering system). V_HHs characteristically have Phe or Tyr, Glu or Gln, Arg or Cys and Gly, Ser, Leu or Phe at positions 37, 44, 45, and 47, respectively, whereas V_Hs have Val, Gly, Leu, and Trp at these four positions.
10. To grow colonies on the stock plate, streak out a frozen stock of TG1 on a minimal plate *(52)* supplemented with thiamine. Incubate at 37 °C for at least 24 h. Seal the plate with parafilm and store at 4 °C for up to a month. It is recommended to grow the TG1 cells on minimal media to ensure that the F pilus, which mediates phage infection, is maintained on the cells. Thiamine is added to the media since TG1 cells are auxotrophic for this vitamin.
11. Skim milk has biotin which may interfere with the binding of biotinylated peptide to streptavidin. Thus, it is recommended to debiotinylate the milk *(59)* or use commercially available biotin-free skim milk.
12. Since peptides are not efficient in coating microtiter plate wells, an indirect coating involving capture of the biotinylated peptide by streptavidin is carried out. Direct coating through covalent linkages is possible, but this may significantly alter the conformation of the peptide resulting in the isolation of antibodies which may only recognize the altered conformation.
13. Binding of biotin pulls the peptide into the streptavidin-binding cavity to a depth of several Angstroms. Thus, to ensure complete display of the peptide on the surface of streptavidin a linker of approximately four amino acids is required between the target peptide and the biotin label. In the case of short peptides, the lack of a linker could lead to the peptide being presented in

the context of streptavidin, resulting in the isolation of binders to streptavidin–peptide complex and not the target peptide.

14. To further accentuate the selection of peptide binders over streptavidin binders, an excess amount of streptavidin may be included in the solution phase.
15. Do not incubate with triethylamine beyond 10 min as phages lose their infectivity significantly.
16. Always keep the phage on ice to prevent possible enzymatic cleavage of gIIIp, due to protease contamination. gIIIp cleavage results in loss of infectivity as well as the antigen-binding activity of the phage particles.
17. Decisions regarding the number of rounds of panning very much depend on the sequence heterogeneity of individual colonies after third or fourth round of panning. Panning may be stopped when sequencing of randomly picked clones reveals that many have the same sequence and when there is good enrichment in terms of the ratio of output to input phage (more than 1,000-fold).
18. In our experience, phages maintain their activity at 4 °C for at least 4 weeks.
19. Two FR1-specific primers were used to construct the V_HH library (*see* **Subheading 3.1**). Depending on which primer a V_HH is derived from, either primerVHBbs or VHBbs1 is used for sub-cloning.
20. Alternatively, commercially available anti-His antibodies can be used for detection. We find that the V_HHs are typically located in the "shock" and/or "sucrose" fractions.
21. For V_HHs with the intrinsic ability to bind to protein A *(60)*, a one-step purification can also be performed on commercially available protein A affinity columns.
22. The data for the binding of the two V_HHs to streptavidin-captured peptide showed good fitting to a steady-state affinity model (**Fig. 20.4**). To confirm that the V_HHs recognized free peptide and not the peptide in complex with streptavidin, data were also collected for the binding of free peptide to immobilized V_HHs. The data for the binding of the V_HHs to captured peptide and for free peptide binding to immobilized V_HHs were in good agreement (data not shown) indicating that both V_HHs recognized a free peptide epitope. The relatively high K_Ds for the interaction of the two V_HHs with peptide suggest that they are not derived from heavy chain antibodies generated as a result of an immune response to the peptide antigen. The affinities of immune heavy chain antibodies are typically in the low nanomolar range.

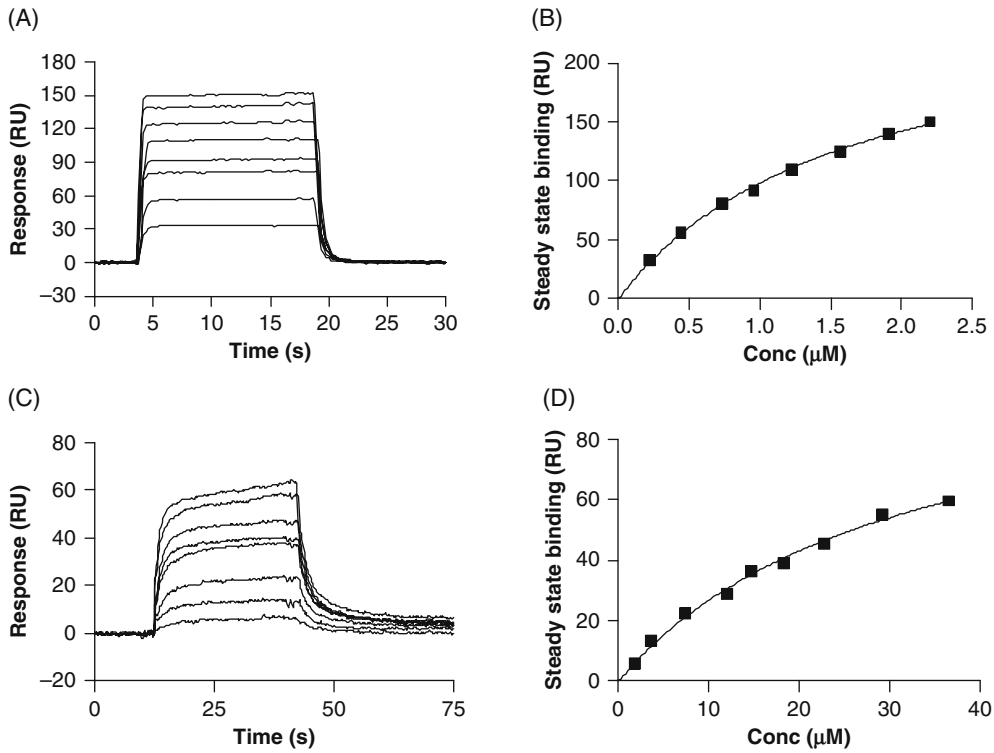


Fig. 20.4. SPR analysis of the binding of FMDV22-1 and FMDV22-2 V_HHs to FMDV22 peptide. **A**, Sensorgram overlay showing FMDV22-1 binding to FMDV22 peptide captured on streptavidin (1,000 RUs) at the concentrations of 0.22, 0.44, 0.73, 0.96, 1.2, 1.6, 1.9, and 2.2 μ M. **B**, Steady-state affinity fitting of the corresponding data in (A). **C**, Sensorgram overlay showing FMDV22-2 binding to FMDV22 peptide captured on streptavidin (1,000 RUs) at the concentrations of 1.8, 3.7, 7.3, 12, 15, 18, 23, and 37 μ M. **D**, Steady-state affinity fitting of the corresponding data in (C). The dissociation constants for the interaction of peptide with FMDV22-1 and FMDV22-2 were determined to be 2 and 30 μ M, respectively.

Acknowledgements

We thank Klaus Nielsen and Lulin Li for providing the peptide antigen and performing the llama immunizations. The assistance of Ginette Dubuc and Shenghua Li with library construction and panning and protein expression is gratefully acknowledged. We thank Tomoko Hirama for performing the SPR analyses.

References

1. Cwirla, S. E., Peters, E. A., Barrett, R. W., & Dower, W. J. (1990). Peptides on phage: a vast library of peptides for identifying ligands. *Proc. Natl. Acad. Sci. U. S. A* **87**, 6378–6382.
2. Scott, J. K. & Smith, G. P. (1990). Searching for peptide ligands with an epitope library. *Science* **249**, 386–390.
3. Smith, G. P. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **228**, 1315–1317.
4. Arbabi-Ghahroudi, M., Desmyter, A., Wyns, L., Hamers, R., & Muyldermans, S. (1997). Selection and identification of single domain antibody fragments from camel heavy-chain antibodies. *FEBS Lett* **414**, 521–526.
5. Barbas, C. F., III, Kang, A. S., Lerner, R. A., & Benkovic, S. J. (1991). Assembly of com-

binatorial antibody libraries on phage surfaces: the gene III site. *Proc. Natl. Acad. Sci. U. S. A* **88**, 7978–7982.
6. Bradbury, A. & Cattaneo, A. (1995). The use of phage display in neurobiology. *Trends Neurosci.* **18**, 243–249.
7. Bradbury, A. (2003). scFvs and beyond. *Drug Discov. Today* **8**, 737–739.
8. Breitling, F., Dubel, S., Seehaus, T., Klewinghaus, I., & Little, M. (1991). A surface expression vector for antibody screening. *Gene* **104**, 147–153.
9. Clackson, T., Hoogenboom, H. R., Griffiths, A. D., & Winter, G. (1991). Making antibody fragments using phage display libraries. *Nature* **352**, 624–628.
10. Davies, J. & Riechmann, L. (1996). Single antibody domains as small recognition units: design and in vitro antigen selection of camelized, human VH domains with improved protein stability. *Protein Eng* **9**, 531–537.
11. Hoogenboom, H. R., Griffiths, A. D., Johnson, K. S., Chiswell, D. J., Hudson, P., & Winter, G. (1991). Multi-subunit proteins on the surface of filamentous phage: methodologies for displaying antibody (Fab) heavy and light chains. *Nucleic Acids Res.* **19**, 4133–4137.
12. Hoogenboom, H. R., de Bruine, A. P., Hufton, S. E., Hoet, R. M., Arends, J. W., & Roovers, R. C. (1998). Antibody phage display technology and its applications. *Immunotechnology.* **4**, 1–20.
13. Lowman, H. B. (1997). Bacteriophage display and discovery of peptide leads for drug development. *Annu. Rev. Biophys. Biomol. Struct.* **26**, 401–424.
14. Marks, J. D., Hoogenboom, H. R., Bonnert, T. P., McCafferty, J., Griffiths, A. D., & Winter, G. (1991). By-passing immunization. Human antibodies from V-gene libraries displayed on phage. *J Mol Biol* **222**, 581–597.
15. McCafferty, J., Griffiths, A. D., Winter, G., & Chiswell, D. J. (1990). Phage antibodies: filamentous phage displaying antibody variable domains. *Nature* **348**, 552–554.
16. Tanha, J., Dubuc, G., Hirama, T., Narang, S. A., & MacKenzie, C. R. (2002). Selection by phage display of llama conventional V(H) fragments with heavy chain antibody V(H)H properties. *J. Immunol. Methods* **263**, 97–109.
17. Marks, J. D. & Bradbury, A. (2004). Selection of human antibodies from phage display libraries. *Methods Mol. Biol.* **248**, 161–176.
18. Sblattero, D. & Bradbury, A. (2000). Exploiting recombination in single bacteria to make large phage antibody libraries. *Nat. Biotechnol.* **18**, 75–80.
19. Winter, G., Griffiths, A. D., Hawkins, R. E., & Hoogenboom, H. R. (1994). Making antibodies by phage display technology. *Annu. Rev Immunol* **12**, 433–455.
20. Felici, F., Castagnoli, L., Musacchio, A., Jappelli, R., & Cesareni, G. (1991). Selection of antibody ligands from a large library of oligopeptides expressed on a multivalent exposition vector. *J Mol. Biol.* **222**, 301–310.
21. Kay, B. K., Adey, N. B., He, Y. S., Manfredi, J. P., Mataragnon, A. H., & Fowlkes, D. M. (1993). An M13 phage library displaying random 38-amino-acid peptides as a source of novel sequences with affinity to selected targets. *Gene* **128**, 59–65.
22. Scott, J. K., Loganathan, D., Easley, R. B., Gong, X., & Goldstein, I. J. (1992). A family of concanavalin A-binding peptides from a hexapeptide epitope library. *Proc. Natl. Acad. Sci. U. S. A* **89**, 5398–5402.
23. Burton, D. R., Barbas, C. F., III, Persson, M. A., Koenig, S., Chanock, R. M., & Lerner, R. A. (1991). A large array of human monoclonal antibodies to type 1 human immunodeficiency virus from combinatorial libraries of asymptomatic seropositive individuals. *Proc. Natl. Acad. Sci. U. S. A* **88**, 10134–10137.
24. Gram, H., Marconi, L. A., Barbas, C. F., III, Collet, T. A., Lerner, R. A., & Kang, A. S. (1992). In vitro selection and affinity maturation of antibodies from a naive combinatorial immunoglobulin library. *Proc. Natl. Acad. Sci. U. S. A* **89**, 3576–3580.
25. Griffiths, A. D. (1993). Production of human antibodies using bacteriophage. *Curr. Opin. Immunol* **5**, 263–267.
26. Hoogenboom, H. R., Marks, J. D., Griffiths, A. D., & Winter, G. (1992). Building antibodies from their genes. *Immunol. Rev.* **130**, 41–68.
27. Muruganandam, A., Tanha, J., Narang, S., & Stanimirovic, D. (2002). Selection of phage-displayed llama single-domain antibodies that transmigrate across human blood-brain barrier endothelium. *FASEB J.* **16**, 240–242.
28. Vaughan, T. J., Williams, A. J., Pritchard, K., Osbourn, J. K., Pope, A. R., Earnshaw, J. C., McCafferty, J., Hodits, R. A., Wilton, J., & Johnson, K. S. (1996). Human antibodies with sub-nanomolar affinities isolated from a large non-immunized phage display library. *Nat. Biotechnol.* **14**, 309–314.
29. Knappik, A., Ge, L., Honegger, A., Pack, P., Fischer, M., Wellnhofer, G., Hoess, A., Wolle, J., Pluckthun, A., & Virnekas, B. (2000). Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. *J. Mol. Biol.* **296**, 57–86.
30. Krebs, B., Rauchenberger, R., Reiffert, S., Rothe, C., Tesar, M., Thomassen, E., Cao,

M., Dreier, T., Fischer, D., Hoss, A., Inge, L., Knappik, A., Marget, M., Pack, P., Meng, X. Q., Schier, R., Sohlemann, P., Winter, J., Wolle, J., & Kretzschmar, T. (2001). High-throughput generation and engineering of recombinant human antibodies. *J. Immunol. Methods* **254**, 67–84.

31. Tanha, J., Xu, P., Chen, Z. G., Ni, F., Kaplan, H., Narang, S. A., & MacKenzie, C. R. (2001). Optimal design features of camelized human single-domain antibody libraries. *J. Biol. Chem* **276**, 24774–24780.
32. Hawkins, R. E., Russell, S. J., & Winter, G. (1992). Selection of phage antibodies by binding affinity. Mimicking affinity maturation. *J Mol. Biol.* **226**, 889–896.
33. Lavoie, T. B., Drohan, W. N., & Smith-Gill, S. J. (1992). Experimental analysis by site-directed mutagenesis of somatic mutation effects on affinity and fine specificity in antibodies specific for lysozyme. *J Immunol* **148**, 503–513.
34. Arap, M. A. (2005). Phage display technology: applications and innovations. *Genet. Mol. Biol.* **28**, 1–9. Ref Type: Journal
35. Conrad, U. & Scheller, J. (2005). Considerations on antibody-phage display methodology. *Comb. Chem High Throughput. Screen.* **8**, 117–126.
36. Kirsch, M., Zaman, M., Meier, D., Dubel, S., & Hust, M. (2005). Parameters affecting the display of antibodies on phage. *J Immunol Methods* **301**, 173–185.
37. Griffiths, A. D., Williams, S. C., Hartley, O., Tomlinson, I. M., Waterhouse, P., Crosby, W. L., Kontermann, R. E., Jones, P. T., Low, N. M., Allison, T. J., & et al (1994). Isolation of high affinity human antibodies directly from large synthetic repertoires. *EMBO J* **13**, 3245–3260.
38. Sidhu, S. S., Li, B., Chen, Y., Fellouse, F. A., Eigenbrot, C., & Fuh, G. (2004). Phage-displayed antibody libraries of synthetic heavy chain complementarity determining regions. *J Mol. Biol.* **338**, 299–310.
39. Bradbury, A., Persic, L., Werge, T., & Cattaneo, A. (1993). Use of living columns to select specific phage antibodies. *Biotechnology (N. Y.)* **11**, 1565–1569.
40. Mutuberria, R., Hoogenboom, H. R., van der, L. E., de Bruine, A. P., & Roovers, R. C. (1999). Model systems to study the parameters determining the success of phage antibody selections on complex antigens. *J Immunol Methods* **231**, 65–81.
41. Cai, X. & Garen, A. (1995). Anti-melanoma antibodies from melanoma patients immunized with genetically modified autologous tumor cells: selection of specific antibodies from single-chain Fv fusion phage libraries. *Proc. Natl. Acad. Sci. U. S. A* **92**, 6537–6541.
42. Palmer, D. B., George, A. J., & Ritter, M. A. (1997). Selection of antibodies to cell surface determinants on mouse thymic epithelial cells using a phage display library. *Immunology* **91**, 473–478.
43. Becerril, B., Poul, M. A., & Marks, J. D. (1999). Toward selection of internalizing antibodies from phage libraries. *Biochem. Biophys. Res. Commun.* **255**, 386–393.
44. Poul, M. A., Becerril, B., Nielsen, U. B., Morisson, P., & Marks, J. D. (2000). Selection of tumor-specific internalizing human antibodies from phage libraries. *J Mol. Biol.* **301**, 1149–1161.
45. Baek, H., Suk, K. H., Kim, Y. H., & Cha, S. (2002). An improved helper phage system for efficient isolation of specific antibody molecules in phage display. *Nucleic Acids Res.* **30**, e18.
46. Chames, P. & Baty, D. (2000). Antibody engineering and its applications in tumor targeting and intracellular immunization. *FEMS Microbiol. Lett.* **189**, 1–8. Ref Type: Journal
47. Harrison, J. L., Williams, S. C., Winter, G., & Nissim, A. (1996). Screening of phage antibody libraries. *Methods Enzymol.* **267**, 83–109.
48. Duenas, M., Malmborg, A. C., Casalvilla, R., Ohlin, M., & Borrebaeck, C. A. (1996). Selection of phage displayed antibodies based on kinetic constants. *Mol. Immunol* **33**, 279–285.
49. Mancini, N., Carletti, S., Perotti, M., Canducci, F., Mammarella, M., Sampaolo, M., & Burioni, R. (2004). Phage display for the production of human monoclonal antibodies against human pathogens. *New Microbiol.* **27**, 315–328
50. Muyldermans, S., Cambillau, C., & Wyns, L. (2001). Recognition of antigens by single-domain antibody fragments: the superfluous luxury of paired domains. *Trends Biochem. Sci.* **26**, 230–235.
51. Tanha, J., Muruganandam, A., & Stanimirovic, D. (2003). Phage Display Technology for Identifying Specific Antigens on Brain Endothelial Cells. *Methods Mol. Med.* **89**, 435–450.
52. Sambrook, J., Fritsch, E. F., & Maniatis, T. (1989). *Molecular Cloning: A laboratory Manual, 2*nd *Ed.* Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
53. Tung, W. L. & Chow, K. C. (1995). A modified medium for efficient electrotransformation of E. coli. *Trends Genet.* **11**, 128–129.
54. Harmsen, M. M., Ruuls, R. C., Nijman, I. J., Niewold, T. A., Frenken, L. G. J., & de Geus, B. (2000). Llama heavy-chain V regions consist of at least four distinct subfamilies

revealing novel sequence features. *Mol. Immunol* **37**, 579–590.
55. Anand, N. N., Dubuc, G., Phipps, J., MacKenzie, C. R., Sadowska, J., Young, N. M., Bundle, D. R., & Narang, S. A. (1991). Synthesis and expression in Escherichia coli of cistronic DNA encoding an antibody fragment specific for a Salmonella serotype B O-antigen. *Gene* **100**, 39–44.
56. MacKenzie, C. R., Sharma, V., Brummell, D., Bilous, D., Dubuc, G., Sadowska, J., Young, N. M., Bundle, D. R., & Narang, S. A. (1994). Effect of C lambda-C kappa domain switching on Fab activity and yield in Escherichia coli: synthesis and expression of genes encoding two anti-carbohydrate Fabs. *Biotechnology N. Y.* **12**, 390–395.
57. Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680–685.
58. Pace, C. N., Vajdos, F., Fee, L., Grimsley, G., & Gray, T. (1995). How to measure and predict the molar absorption coefficient of a protein. *Protein Sci.* **4**, 2411–2423.
59. Yau, K. Y., Dubuc, G., Li, S., Hirama, T., MacKenzie, C. R., Jermutus, L., Hall, J. C., & Tanha, J. (2005). Affinity maturation of a V(H)H by mutational hotspot randomization. *J Immunol Methods* **297**, 213–224.
60. Spinelli, S., Frenken, L., Bourgeois, D., de Ron, L., Bos, W., Verrips, T., Anguille, C., Cambillau, C., & Tegoni, M. (1996). The crystal structure of a llama heavy chain variable domain. *Nat. Struct. Biol.* **3**, 752–757.

Chapter 21

Internet Resources of Interest to Bacteriophage Workers

Andrew M. Kropinski

Abstract

The Internet provides a myriad of useful tools for the phage worker including access to culture collections, specific databases, tools for gene identification, and whole genome comparisons, lecture notes, information on upcoming scientific meetings, books, etc.

Key words: ASM, American Society for Microbiology, BEG, Bacteriophage Ecology Group, EM, Database, NCBI, EMBL-EBI, Pittsburgh Bacteriophage Institute, T4, taxonomy, nomenclature, ATCC, DSMZ, Felix d'Herelle Reference Centre, books, meetings, companies.

1 Introduction

The Internet is simultaneously an incredible and a frustrating resource for all who use it. Sites have the unfortunate habit of moving servers, being renamed or being deleted. As a result, I have been very conservative with the sites (URLs) listed, and these should not be considered, by any means, the only sites that are available. An updated list of sites will be maintained by the authors.

2 Internet Sites

2.1 Good places to start

1. American Society for Microbiology Division M: Bacteriophage: http://www.asm.org/division/m/M.html
2. www.phage.org—The Bacteriophage Ecology Group Home of Phage Ecology and Evolutionary Biology (S.T. Abedon): http://www.mansfield.ohio-state.edu/~sabedon/

Martha R. J. Clokie, Andrew M. Kropinski (eds.), *Bacteriophages: Methods and Protocols, Volume 2: Molecular and Applied Aspects, vol. 502*, © 2009 Humana Press, a part of Springer Science+Business Media
DOI 10.1007/978-1-60327-565-1_21 Springerprotocols.com

3. Phage: http://www.sci.sdsu.edu/~smaloy/MicrobialGenetics/topics/phage/
4. All the Virology on the WWW: http://www.virology.net/

2.2 Pictures of phage

1. Bacteriophage Ecology Group Phage Images: select images at http://www.mansfield.ohio-state.edu/~sabedon/
2. ASM Division M: Index of phage pages: http://www.asm.org/division/m/smile.html

2.3 Phage structure

1. VIPERdb: Virus Particle Explorer: http://viperdb.scripps.edu/
2. Virus Ultrastructure: http://web.uct.ac.za/depts/mmi/stannard/linda.html

2.4 Phage Genomes in DNA Databases:

1. NCBI Phage Genomes: http://www.ncbi.nlm.nih.gov/genomes/genlist.cgi?taxid=10239&type=6&name=Phages
2. NCBI Complete Microbial Genomes (please note that most bacteria are lysogenic):http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi?view=1

   These sequences can be downloaded in a variety of formats from the GenBank genome ftp site (ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/). These include in alphabetical order: *.asn (ASN.1 file, print format), *.faa (FASTA formatted protein sequences), *.ffn (FASTA formatted open reading frames), *.fna (FASTA nucleic acid file), *.gbk (GenBank flat file format), *.gbs (GenBank summary file format), *.ptt (Protein table), *.tab (table to assemble genome), and *.val (ASN.1 binary format).
   Please note that annotated prophage genomes can be downloaded from NCBI (search: Genomes) by accessing the host genome sequence, selecting the end points of the prophage in the "Range: from Begin to End" clicking on "Refresh," and saving the resulting mini-gbk file. It has been my experience that prophages are generally poorly annotated.
3. NCBI Draft Bacterial Genome Assembly Sequences: http://www.ncbi.nlm.nih.gov/genomes/genlist.cgi?taxid=2&type=3&name=Bacterial%20Assembly%20Sequences
4. NCBI Draft Archael Genome Assembly Sequences: http://www.ncbi.nlm.nih.gov/genomes/genlist.cgi?taxid=2157&type=3&name=Archaea%20Assembly%20Sequences
5. EMBL-EBI Genome: Phage: http://www.ebi.ac.uk/genomes/phage.html
6. EMBL-EBI Genome: Archaea: http://www.ebi.ac.uk/genomes/archaea.html
7. EMBL-EBI Genome: Bacteria: http://www.ebi.ac.uk/genomes/bacteria.html
8. Pittsburgh Bacteriophage Institute: http://www.pitt.edu/~biology/Dept/Frame/pbi.htm

9. Genomes of the T4-like Phages: http://phage.bioc.tulane.edu/
10. Prophage Database *(1)* (http://bicmku.in:8082/prophagedb/ or http://ispc.weizmann.ac.il/prophagedb) and Prophage Finder *(2)* at http://bioinformatics.uwp.edu/~phage/ProphageFinder.php
11. ACLAME (A CLAssification of genetic Mobile Elements) contains databases on phages and prophages: http://aclame.ulb.ac.be/

2.5 Taxonomy and nomenclature

1. NCBI Taxonomy Browser (The NCBI Taxonomy Homepage): http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/
2. ICTV Taxonomy and Index to Virus Classification and Nomenclature Taxonomic lists and catalog of viruses: http://www.ncbi.nlm.nih.gov/ICTVdb/Ictv/fr-index.htm
3. Naming bacteriophages (Hans-Wolfgang Ackermann & Stephen Tobias Abedon). Bacteriophage Ecology Group (BEG) News *July 1, 2001 issue (volume 9)*: http://www.mansfield.ohio-state.edu/~sabedon/bgnws009.htm
4. Hans-Wolfgang Ackermann & Stephen Tobias Abedon (2001). Bacteriophage Names 2000. *The Bacteriophage Ecology Group*. http://www.mansfield.ohio-state.edu/~sabedon/names.htm

2.6 Major phage culture collections

1. American type culture collection (ATCC)
   P.O. Box 1549
   Manassas, VA 20108
   USA
   Telephone number: (800) 638-6597
   URL (general): http://www.atcc.org/
   URL (phage): http://www.atcc.org/ATCCAdvancedCatalogSearch/tabid/112/Default.aspx
2. DSMZ_DSMZ_Deutsche Sammlung von Mikroorganismen und ZelIkulturen GmbH und ZelIkulturen GmbH
   Mascheroder Weg 1b 38124 Braunschweig Germany
   Telephone number: +49-531-2616-0
   URL (general): http://www.dsmz.de
3. Félix d'Hérelle Reference Center for Bacterial Viruses
   Dr. Sylvain Moineaux
   Université Laval
   Département de Biochimie et de Microbiologie
   Québec, QC, G1K 7P4
   Canada
   Telephone number: 418-656-3712
   URL: http://www.phage.ulaval.ca/index.php

2.7 Partial list of books on phage published in the last 10 years (for a complete list, see: Stephen Abedon's Internet review at http://en.wikipedia.org/wiki/Phage_monographs)

1. The Bacteriophages, 2nd edition (Editor: R. Calendar). 2006. Oxford University Press, New York. ISBN 0-19-514850-9.
2. Phages: Their Role in Bacterial Pathogenesis and Biotechnology (Editors: M.K. Waldor, D.I. Friedman, & S.L. Adhya). 2005. ASM Press, Washington, D.C. ISBN 1-55581-307-0.
3. Viral Genome Packaging Machines: Genetics, Structure, and Mechanisms (Editor: C.E. Catalano). 2005. Kluwer Academic/Plenum Publishers, New York. ISBN 0-306-48227-4.
4. Bacteriophages: Biology and Applications (Editors: E. Kutter & A. Sulakvelidze). 2004. CRC Press, ISBN 0-84931-336-8.
5. A Genetic Switch (3rd edition). 2004. M. Patshne. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York. ISBN 0-87969-716-4.
6. Gesund Durch Viren—Ein Ausweg Aus Der Antibiotika-Krise (Healthy Through Viruses—a way out of the antibiotic-resistance crisis.). 2003. T. Häusler. Piper, München, Germany. ISBN 3-49204-520-0.
7. Bacterial and Bacteriophage Genetics. 2006. E.A. Birge. Springer-Verlag, New York. ISBN 0-387-23919-7
8. Viruses vs Superbugs: a Solution to the Antibiotic Crisis. 2006. T. Häusler. Pargrave-Macmillan Science, ISNB 1403987645.
9. Félix D'Herelle and the Origins of Molecular Biology. 1999. W.C. Summers, Yale University Press, Cumberland, RI. ISBN 0-30007-127-2

2.8 Phage meetings

1. Biennial International Evergreen Phage Biology Meeting (Evergreen State College, Olympia, WA, USA): http://academic.evergreen.edu/projects/phage/generalmeetingcalendar.htm
2. Molecular Genetics of Bacteria and Phages Meetings: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York; or University of Wisconsin-Madison.
3. Biennial Conference on Phage/Virus Assembly: Location—Variable
4. American Society for Microbiology Division M: http://www.asm.org/division/m/M.html
5. International Congress of Virology (International Union of Microbiological Societies): http://www.iums.org/
6. Phage meetings: http://en.wikipedia.org/wiki/Phage_meetings

2.9 Companies involved in phage, and phage therapy research

1. Biophage Pharma, Inc. (Montreal, QC, Canada): http://www.biophagepharma.net/index.html
2. Exponential Biotherapies, Inc. (Washington, DC, U.S.A.): http://www.expobio.com/
3. Gangagen Biotechnologies Pvt Ltd (Palo Alto, CA, U.S.A.): http://www.gangagen.com/

4. Hexal Genentech (Holzkirchen, Germany): http://www.hexal-gentech.de/
5. Intralytix, Inc. (Baltimore, MD, U.S.A.): http://www.intralytix.com/
6. New Horizons Diagnostics, Inc. (Columbia, MD, U.S.A.): http://www.nhdiag.com/index.htm
7. Novolytics Ltd. (Coventry, United Kingdom): http://www.novolytics.co.uk/about_us.html
8. Phage Biotech Ltd. (Rehovot, Israel): http://www.phage-biotech.com/
9. PhageInternational, Inc. (Los Altos, CA, U.S.A.): http://www.phageinternational.com/
10. Phage Therapy Center (Tbilisi, Georgia): http://www.phagetherapycenter.com/pii/PatientServlet?command=static_home&secnavpos= −1&language=0
11. Targanta Therapeutics, Inc. (St. Laurent, QC, Canada): http://www.targanta.com/

Many of the sites listed above have useful information on bacteriophages therapy (phagotherapy): Bacteriophage Ecology Group Phage Therapy References http://www.mansfield.ohio-state.edu/~sabedon/(listed under links); and, Elizabeth Kutter's (Evergreen State College, Olympia, WA), site on "Phage Therapy" at http://www.evergreen.edu/phage/phagetherapy/phagetherapy.htm.

2.10 Additional phage methods

1. Ebioinfogen Phage -related Protocols: http://www.ebioinfogen.com/phage.htm
2. Favorite King Lab Recipes & Protocols (Jonathan King, MIT): http://web.mit.edu/king-lab/www/cookbook/cookbook.htm
3. Protocol Online: http://www.protocol-online.org/prot/Molecular_Biology/Phage/index.html
4. Protocols from the Thomas Lab (George J. Thomas, University of Missouri-Kansas City): http://sbs.umkc.edu/gjthomas-lab/protocols/index.html
5. The Guide to Phage Genomics: Isolation, Purification, Cloning, Sequencing, Assembling, Annotating and Analyzing Phage Genomes: http://www.sci.sdsu.edu/PHAGE/guide.html

2.11 Lectures on bacteriophages

1. Gene Meyer (University of South Carolina, School of Medicine): http://pathmicro.med.sc.edu/mayer/phage.htm
2. Martin E. Mulligan (Department of Biochemistry, Memorial University, Canada):
   a. http://www.mun.ca/biochem/courses/3107/Lectures/Topics/bacteriophage.html
   b. http://www.mun.ca/biochem/courses/3107/Lectures/Topics/bacteriophage_replication.html

c. http://www.mun.ca/biochem/courses/4103/topics/Lambda/Lambda.html
d. http://www.mun.ca/biochem/courses/4103/topics/Lambda/Lambda_immunity.html

3. Phage reading: on Nobel Prize winners and their page-related research http://www.drjreid.com/phage_biology_url.htm

References

1. Srividhya, K.V., R.V. Greeta, L. Raghavenderan, M. Preeti, J. Prilusky, M. Sankarnarayanan, J.L. Sussman, and S. Krishnaswamy. 2006. Database and comparative identification of prophages, p. 863–868. *In* International Conference on Intelligent Computing 2006. Springer-Verlag, Berlin.
2. Bose, M. and R. Barber. 2006. Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. In Silico Biology *6*:0020.

INDEX

G

H

I

L

M

N

O

P

R

S

T

U

V

W

Y

INDEX

G

H

I

L

M

N

O

P

R

S

T

U

V

W

Y