Springer Protocols

Methods in Molecular Biology 551

Molecular Epidemiology of Microorganisms

Methods and Protocols

Edited by

Dominique A. Caugant

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

Molecular Epidemiology of Microorganisms

Edited by

Dominique A. Caugant

Norwegian Institute of Public Health and University of Oslo, Oslo, Norway

Humana Press

Editor
Dominique A. Caugant
Norwegian Institute of Public Health and University of Oslo
Oslo, Norway
Dominique.Caugant@fhi.no

ISBN: 978-1-60327-998-7 e-ISBN: 978-1-60327-999-4
ISSN: 1064-3745 e-ISSN: 1940-6029
DOI: 10.1007/978-1-60327-999-4
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009920074

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The development and rapid implementation of molecular genotyping methods have revolutionized the possibility for differentiation and classification of microorganisms at the subspecies level. Investigation of the species diversity is required to determine molecular relatedness of isolates for epidemiological studies. Methods for molecular epidemiology of microorganisms must be highly reproducible and provide effective discrimination of epidemiologically unrelated strains.

A wide range of techniques has been applied to the investigation of outbreaks of transmissible disease, and these have been critical in unraveling the route of spread of pathogens for humans, animals, and plants. The choice of a molecular method will depend on the type of questions to be addressed, on the degree of genetic diversity of the species to be analyzed, and on the mechanisms responsible for generation of the diversity. The applications of molecular methods, singly or in combination, have greatly contributed in the past two decades to basic microbial science and public health control strategies.

Molecular Epidemiology of Microorganisms: Methods and Protocols brings together a series of methods-based chapters with examples of application to some of the most important microbes. Both traditional and novel techniques are described, and the type of information that can be expected to be obtained by their application is indicated.

I am indebted to all internationally respected colleagues who have provided state-of-the-art chapters for inclusion in this book. I am very grateful for their outstanding contributions, enthusiasm for the project, and friendship. I would like to thank John Walker at Humana Press for the invitation to put this book together and his continuous encouragement.

Dominique A. Caugant

Contents

Contributors

David Michael Aanensen • *Department of Infectious Disease Epidemiology, Imperial College London, London, UK*

Pablo Bifani • *Pasteur Institute, Scientific Institute of Public Health, Belgium*

Lawrence B. Blyn • *Ibis Biosciences, Carlsbad, CA, USA*

Jon S. Brazier • *Anaerobe Reference Laboratory, NPHS Microbiology Cardiff, University Hospital of Wales, Heath Park, Cardiff, UK*

Eleanor M. Cottam • *Institute for Animal Health, Pirbright, Surrey, UK*

Jeffrey R. Driscoll • *Division of Tuberculosis Elimination, Centers for Disease Control and Prevention, Atlanta, GA, USA*

David J. Ecker • *Ibis Biosciences, Carlsbad, CA, USA*

Mark W. Eshoo • *Ibis Biosciences, Carlsbad, CA, USA*

Tamara Feldblyum • *Food and Drug Administration, Center for Devices and Radiological Health, Office of In Vitro Diagnostic Devices Evaluation and Safety, Rockville, MD, USA*

Alexander W. Friedrich • *Institute for Hygiene, University of Münster, Münster, Germany*

Norman K. Fry • *Respiratory and Systemic Infection Laboratory, Health Protection Agency, Centre for Infections, London, UK*

Ibtissem Grissa • *Department of Genetics and Microbiology, University of Paris XI, Orsay, France*

Thomas A. Hall • *Ibis Biosciences, Carlsbad, CA, USA*

Marie Hallin • *Laboratoire de Référence MRSA-Staphylocoques, Department of Microbiology, Hôpital Erasme, Université Libre de Bruxelles, Brussels, Belgium*

William Paul Hanage • *Department of Infectious Disease Epidemiology, Imperial College London, London, UK*

James C. Hannis • *Ibis Biosciences, Carlsbad, CA, USA*

John P. Hays • *Erasmus MC, University Hospital Rotterdam, Department of Medical Microbiology and Infectious Diseases, Unit Research and Development, Rotterdam, The Netherlands*

Kelli L. Hiett • *Poultry Microbiological Safety Research Unit, Russell Research Center, Agricultural Research Service, U.S. Department of Agriculture, Athens, GA, USA*

Steven A. Hofstadler • *Ibis Biosciences, Carlsbad, CA, USA*

Carol Holm-Hansen • *Division of Infectious Disease Control, Norwegian Institute of Public Health, Oslo, Norway*

Stephan Huehn • *Federal Institute for Risk Assessment, National Salmonella Reference Laboratory, Berlin, Germany*

Ana Belén Ibarz Pavón • *Department of Zoology and Peter Medawar Building for Pathogen Research, University of Oxford, Oxford, UK*

Timothy J. Johnson • *Department of Veterinary and Biomedical Sciences, University of Minnesota, St. Paul, MN, USA*
Keith A. Jolley • *Department of Zoology and Peter Medawar Building for Pathogen Research, University of Oxford, Oxford, UK*
Donald P. King • *Institute for Animal Health, Pirbright, Surrey, UK*
Nick J. Knowles • *Institute for Animal Health, Pirbright, Surrey, UK*
Barry Kreiswirth • *Tuberculosis Centre, Public Health Research Institute, Newark, NJ, USA*
Ed J. Kuijper • *Reference Laboratory for Clostridium difficile, Medical Microbiology Department, LUMC, Leiden, and The National Institute for Public Health and Environment, Bilthoven, The Netherlands*
Natalia Kurepina • *Tuberculosis Centre, Public Health Research Institute, Newark, NJ, USA*
Martin C. J. Maiden • *Department of Zoology and Peter Medawar Building for Pathogen Research, University of Oxford, Oxford, UK*
Burkhard Malorny • *Federal Institute for Risk Assessment, National Salmonella Reference Laboratory, Berlin, Germany*
Christian Massire • *Ibis Biosciences, Carlsbad, CA, USA*
Barun Mathema • *Tuberculosis Centre, Public Health Research Institute, Newark, NJ, USA*
Lisa K. Nolan • *Department of Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, IA, USA*
Tansy M. Peters • *Health Protection Agency, Centre for Infections, London, UK*
Christine Pourcel • *Department of Genetics and Microbiology, University of Paris XI, Orsay, France*
Rangarajan Sampath • *Ibis Biosciences, Carlsbad, CA, USA*
Paul H. M. Savelkoul • *Department of Medical Microbiology and Infection Control, VU University Medical Center, Amsterdam, The Netherlands*
Bruce S. Seal • *Poultry Microbiological Safety Research Unit, Russell Research Center, Agricultural Research Service, U.S. Department of Agriculture, Athens, GA, USA*
Marc J. Struelens • *Laboratoire de Référence MRSA-Staphylocoques, Department of Microbiology, Hôpital Erasme, Université Libre de Bruxelles, Brussels, Belgium*
Hervé Tettelin • *Institute for Genome Sciences, Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, MD, USA*
Michel Tibayrenc • *IRD Representative Office, French Embassy, Bangkok, Thailand*
Kirsti Vainio • *Division of Infectious Disease Control, Norwegian Institute of Public Health, Oslo, Norway*
Alex van Belkum • *Erasmus MC, University Hospital Rotterdam, Department of Medical Microbiology and Infectious Diseases, Unit Research and Development, Rotterdam, The Netherlands*
Renate J. van den Berg • *Anaerobe Reference Laboratory, NPHS Microbiology Cardiff, University Hospital of Wales, Heath Park, Cardiff, UK*
Elisabeth van Pelt-Verkuil • *Hogeschool Leiden, Department of Post Graduate Courses, Leiden, The Netherlands*

Gilles Vergnaud • *DGA/D4S–Mission pour la Recherche et l'Innovation Scientifique (MRIS), Armées, and Department of Genetics and Microbiology, University of Paris XI, Orsay, France*
Paolo Visca • *Department of Biology, University Roma Tre, and Molecular Microbiology Unit, National Institute for Infectious Diseases "Lazzaro Spallanzani," Rome, Italy*
Jemma Wadsworth • *Institute for Animal Health, Pirbright, Surrey, UK*
Xiao-Ming Wang • *Pasteur Institute, Scientific Institute of Public Health, Belgium*

Chapter 10

Spoligotyping for Molecular Epidemiology of the *Mycobacterium tuberculosis* Complex

Jeffrey R. Driscoll

Abstract

Spacer oligonucleotide typing, or spoligotyping, is a rapid, polymerase chain reaction (PCR)-based method for genotyping strains of the *Mycobacterium tuberculosis* complex (MTB). Spoligotyping data can be represented in absolute terms (digitally), and the results can be readily shared among laboratories, thereby enabling the creation of large international databases. Since the spoligotype assay was standardized more than 10 yr ago, tens of thousands of isolates have been analyzed, giving a global picture of MTB strain diversity. The method is highly reproducible and has been developed into a high-throughput assay for large molecular epidemiology projects. In the United States, spoligotyping is employed on nearly all newly identified culture-positive cases of tuberculosis as part of a national genotyping program. The strengths of this method include its low cost, its digital data results, the good correlation of its results with other genetics markers, its fair level of overall differentiation of strains, its high-throughput capacity, and its ability to provide species information. However, the method's weaknesses include the inability of spoligotyping to differentiate well within large strain families such as the Beijing family, the potential for convergent evolution of patterns, the limited success in improving the assay through expansion, and the difficulty in obtaining the specialized membranes and instrumentation.

Key words: Epidemiology, genotyping, mycobacteria, spoligotyping, tuberculosis.

1. Introduction

DNA fingerprinting or genotyping of *Mycobacterium tuberculosis* complex (MTB) strains became a priority in the United States when in the early 1990s a staggering increase in cases of multidrug-resistant tuberculosis (TB) was observed in New York City *(1)*. Epidemiologists needed to know which cases were linked and where transmissions were occurring. They also needed to determine the size of the outbreak and to try to prevent further

D.A. Caugant (ed.), *Molecular Epidemiology of Microorganisms*, Methods in Molecular Biology, Vol. 551
DOI: 10.1007/978-1-60327-999-4_10,

transmissions. The primary genotyping method available at the time, insertion element (IS) *6110*-based restriction fragment length polymorphism (RFLP) analysis *(2)*, provided excellent differentiation but required specialized software for analysis of the data as well as relatively long turnaround times for reporting of the results. Weeks or months could be required for the level of growth in culture necessary for performance of RFLP analysis. Data analysis required specialized matching software and expert interpretation for relating similar, but not identical, patterns.

Genotyping methods that could employ amplification of nucleic acids were assessed in efforts to develop an alternative to RFLP analysis. The first widely adopted polymerase chain reaction (PCR)-based method for genotyping was spacer oligonucleotide typing or spoligotyping. Kamerbeek et al. *(3)* described a reverse-hybridization protocol to assay for the presence or absence of 43 specific DNA spacer sequences in the direct repeat (DR) region that had been identified in the strains *M. tuberculosis* H37Rv and *Mycobacterium bovis* BCG (**Fig. 1**). The majority of the 43

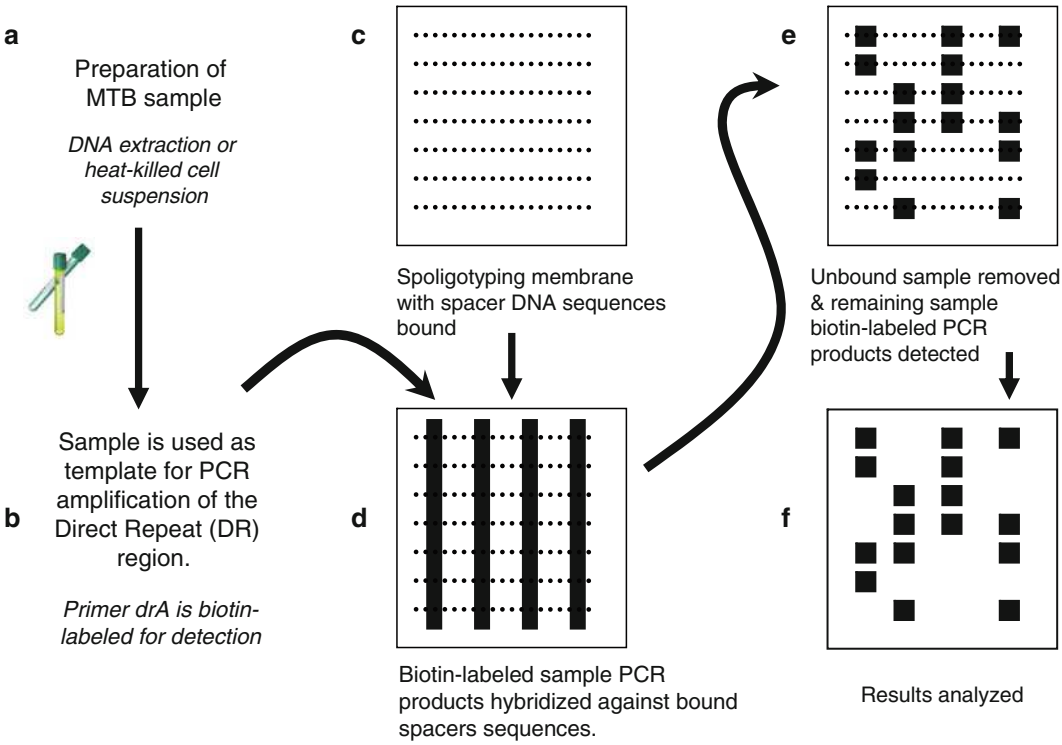


Fig. 1. Basis of the spoligotyping methodology. (**a**) A spoligotyping membrane. Dashed lines indicate the location of the bound polymorphic oligonucleotides, one corresponding to each of the 43 unique spacer sequences utilized in the assay. (**b**) The hybridization of the amplified samples (black bars) against each of the bound oligonuleotides. (**c**) The excess and nonspecifically bound sample is removed through a series of washes, and the remaining bound PCR products from the sample are detected. (**d**) A representation of the final results.

spacers were present in both H37Rv and BCG, but spacers 20, 21, and 33–36 were not present in H37Rv, and spacers 3, 9, 16, and 39–43 were missing in strains of BCG.

The DR region consists of a repeated 36-bp sequence interspersed with nonrepetitive 31- to 41-bp long DNA segments called spacer sequences *(4)*. Spoligotypes evolve through the loss of spacer sequences, presumably through homologous recombination of the DRs and excision of the recombined material during DNA replication. Spacer sequences can also appear to be lost through rearrangements by ISs like IS*6110*. Once spacers are lost, they are not regained, so the evolution is unidirectional. This unidirectional evolution through loss of spacers offers a clear model for evolution but also presents a challenge since a strain's spoligotype can evolve in such a way that it comes to resemble the signature associated with a different spoligotype family. The ability to encode spoligotyping data in a numerical format (**Fig. 2**) *(5)* immediately made the results readily shareable among laboratories and enabled the creation of an international database (SpolDB) *(6)*. This development allowed investigators to survey strain diversity and uncover global strain families, such as the Beijing and the Latin America Mediterranean (LAM) families.

Spoligotyping has been very successful in providing a tool for the rapid acquisition of MTB genotyping information and for the establishment of a global picture of MTB diversity *(6)*.

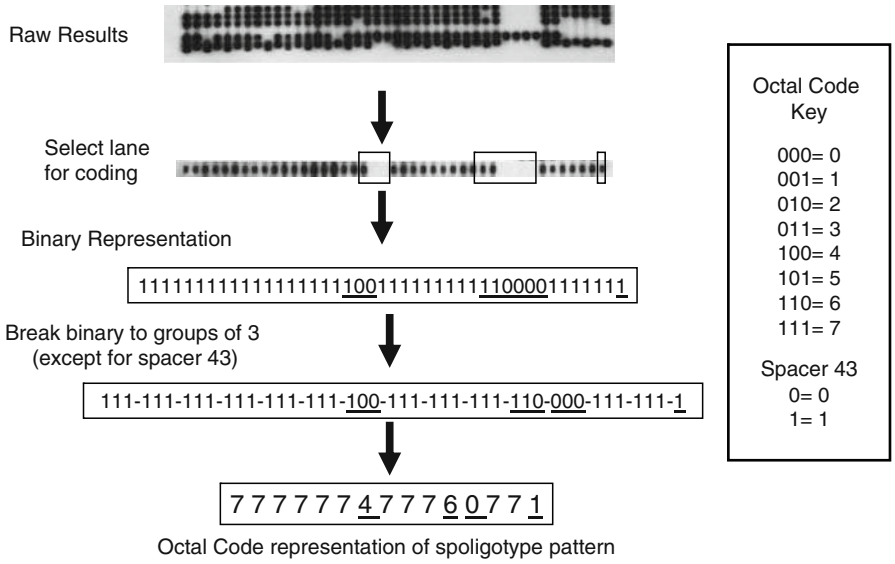


Fig. 2. Conversion of a raw spoligotype results to octal code representation *(5)*. The raw hybridization pattern is converted to a binary representation using 1's (indicating hybridization detected) and 0's (no hybridization detected). The binary string is separated into 14 groups of three, with spacer 43 remaining ungrouped. Each binary triplet is converted to the appropriate octal code designation (shown). The 15th digit of the octal is either 1 or 0 depending solely on the result for spacer 43.

However, the need still exists for supplemental genotyping information. The availability of multiple genotyping markers allows us to be able to "zoom in" to establish specific potential patient-to-patient transmissions and "zoom out" to examine regional and global trends in the spread of tuberculosis strains *(7)*. Alternative DNA fingerprinting methods may supplant spoligotyping in the future if more powerful markers are identified, widely adopted, and their patterns collected into large collaborative databases. For the present, the combination of spoligotype and mycobacterial interspersed repetitive unit (MIRU) data provides a good basis for molecular epidemiology *(8)*, although IS*6110*-based RFLP can still be required in a number of cases for optimal genetic cluster analysis *(9)*.

1.1. Species Identification Within the M. tuberculosis *Complex*

The MTB complex is made up of a group of closely related species: *Mycobacterium africanum*, *M. bovi*s, *Mycobacterium caprae, M. tuberculosis, Mycobacterium microti*, *Mycobacterium canettii,* and *Mycobacterium pinnipedii*. The presence or absence of certain spacer sequences acts as a signature for presumptive species identification *(10)*. For example, *M. bovis* isolates do not hybridize to spacers 39–43 but do generally hybridize to spacers 33–36 *(3)*. *Mycobacterium africanum* isolates do not hybridize to spacers 8, 9, and 39 but do generally hybridize to spacers 33–36. *Mycobacterium microti*, *M. canettii,* and *M. pinnipedii* have very different spoligotype patterns from the members of the MTB complex, which are more associated with human infections. These three species typically hybridize to few if any, in the case of M. Canettii of the traditional 43 spoligotyping spacers *(6,11)*.

1.2. Selective Versus Universal Genotyping

For a public health program, the choice between genotyping only certain MTB strains (selective genotyping) and genotyping every isolate (universal genotyping) comes down to cost issues and the capability to integrate the data to into program activities. The benefits of universal genotyping include earlier identification of false-positive MTB cultures (e.g., due to laboratory cross-contamination), discovery of unsuspected cases of MTB transmission (i.e., linking patients who had not previously been identified as contacts through conventional methods), confirmation of species identification within the MTB complex, and capability to generate a database to examine strain diversity in a particular region for monitoring program success in the control of tuberculosis *(1)*. Universal genotyping enables shorter turnaround times inasmuch as a method like spoligotyping can be performed as a routine activity in the laboratory workup of a patient's MTB strain *(12)*. Selective genotyping, in contrast, can entail requests for analysis of isolates weeks or months after the clinical mycobacteriology laboratory has received the specimen, and retrieval from archival storage may be difficult.

1.3. False Clustering Due to Commonality of Spoligotypes

In areas where particular genetic families are grossly dominant in a TB population, such as Beijing in East Asia *(13)*, spoligotyping without additional genotyping information is of limited value *(8,14)*. The Beijing spoligotype is highly stable, and variants are rarely observed.

Knowledge of the MTB strain diversity in an area is important in establishing the significance of genotyping matches for all fingerprinting methods, but this is especially true for spoligotyping *(15)*. **Table 1** lists the ten most commonly observed spoligotypes in SpolDB4 *(6)*. Generally, a finding that two patients match by one of these spoligotypes does not in of itself prove that the two strains are identical. Additional genotyping data acquired through MIRU or RFLP analysis is required in most cases to establish the significance of a typing match. However, in a well-characterized population, the appearance of two strains with a matching unique spoligotype pattern is likely to be significant, especially if other factors (**Fig. 3**) are present.

Table 1
Ten Most Frequently Observed Spoligotype Patterns in the Fourth International Spoligotype Database (SpolDB4)

Lineage	Spoligotype	Octal Code	SpolDB4 Frequency
T1	●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●◿◿◿◿●●●●●●●	777777777760771	17.85
Beijing	◿◿◿◿◿◿◿◿◿◿◿◿◿◿◿◿◿◿◿◿◿◿◿◿◿◿◿◿◿◿◿◿◿◿●●●●●●●●●	000000000003771	11.28
BOV2	●●◿●●◿●◿◿◿◿◿●●●◿●●●●●●●●●●●●●●●●●●●●●●◿◿◿◿◿	664073777777600	7.71
BOV1	●●◿●●●●●◿●●●●●●◿●●●●●●●●●●●●●●●●●●●●●●◿◿◿◿◿	676773777777600	6.68
Haarlem3	●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●◿●◿◿◿◿●●●●●●●	777777777720771	5.46
LAM9	●●●●●●●●●●●●●●●●●●●●●◿◿◿◿●●●●●●●●◿◿◿◿●●●●●●●	777777607760771	5.37
Haarlem1	●●●●●●●●●●●●●●●●●●●●●●●●●●●◿◿◿◿◿◿●◿◿◿◿●●●●●●●	777777774020771	3.44
CAS1-Delhi	●●●◿◿◿◿●●●●●●●●●●●●●●●◿◿◿◿◿◿◿◿◿◿◿◿●●●●●●●●●	703777740003771	3.00
X2	●●●●●●●●●●●●●●●●●●●◿●●●●●●●●●●●●●●●◿◿◿◿●●◿◿◿◿●	777776777760601	2.98
EAI5	●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●◿◿◿◿●◿●●●●●●●●●	777777777413771	2.82

Source: From **ref. 6.**

For spoligotype hybridization patterns, a closed circle indicates hybridization observed at that spacer sequence, and a gray triangle indicates no hybridization. The spacers 1 through 43 are shown in sequence from left to right.

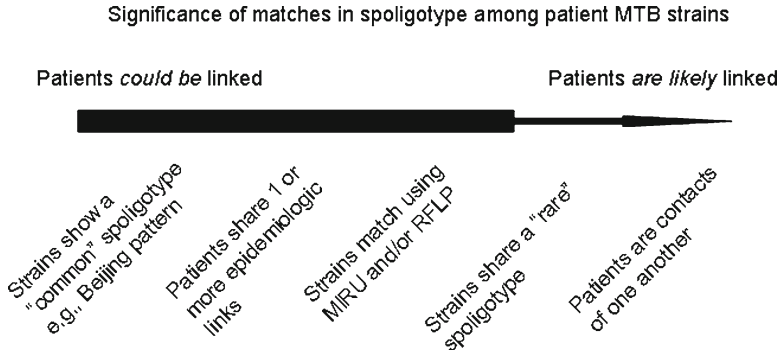


Fig. 3. Factors that aid in the assessment of the significance of a match between spoligotypes. Examples of genotyping and epidemiological data that are useful in deciding whether patients with matching spoligotypes are linked.

1.4. Application to Epidemiology

1.4.1. Application of Spoligotyping to Public Health Programs

With tuberculosis control programs incorporating genotyping data into their routine activities, the need to generate data as early as possible became important in order to direct contact investigations and to identify cases of false-positive cultures (e.g., laboratory cross-contamination) *(16)*. The New York City program was the first large-scale attempt to achieve this *(1)*. The largest genotyping program currently in operation is in the United States. It was developed by the Centers for Disease Control and Prevention, which has also developed a manual for implementation of genotyping data into routine tuberculosis control practice by state and local health departments *(17)*. The ability to obtain spoligotype results from early growth cultures, or even primary specimens, means a "real-time" approach to MTB genotyping is possible *(18)*. However, a single approach to the analysis of genotyping data for every program does not work. A program in an area with a low incidence of TB may find few matches among patient isolates, suggesting a low occurrence of recent transmissions *(19)*. In a high-incidence area a program will encounter a greater number of MTB strains with the more frequently observed spoligotypes (**Table 1**). This obscures the picture of recent transmission versus distant transmission *(1)*, thereby necessitating the use of additional genotyping assays, such as RFLP and MIRU *(18)*.

1.4.2. Identification of False-Positive Cultures of Tuberculosis

Laboratory cross-contamination of patient samples continues to be a problem that results in false diagnoses of tuberculosis. Among the most obvious examples have been laboratory cross-contamination events with the common laboratory strains H37Ra and H37Rv affecting determinations for multiple patient specimens *(17,20,21)*. H37Ra and H37Rv were derived from the same parent isolate, H37, collected in 1905. They share the

same spoligotype and have similar although not identical RFLP patterns *(21)*. To date, there have been no reports of a true patient isolate sharing the H37 spoligotype. Therefore, when a patient isolate is found to have a spoligotype matching H37, a laboratory cross-contamination event is likely to have occurred *(17)*. Cross-contamination of a patient's sample by a different patient's sample or mislabeling errors at the site of collection or the laboratory often requires further investigation, typically involving performing additional genotyping assays and review of patient clinical data *(17)*. It is important to remember that confirmation of a false-positive or cross-contaminated MTB specimen applies solely to the culture results; the diagnosis of tuberculosis in the patient is still made based on the entire clinical presentation.

2. Materials

2.1. Stock Buffers

1. 20X SSPE: 0.2*M* Na_2HPO_4, 3.6*M* $NaCl_2$, 20 m*M* ethylenediaminetetraacetic acid (EDTA), final pH should be 7.4–7.6, autoclaved and stable for 1 yr.
2. 0.5*M* EDTA, pH 8.0, autoclaved and stable for 1 yr.
3. 10% (w/v) sodium dodecyl sulfate (SDS), made fresh as required.

2.2. Polymerase Chain Reaction

Primers for DR region amplification: DRa (GGTTTTGGGTCTGACGAC, 5′ biotinylated) and DRb (CCGAGAGGGGACGGAAAC). Store reconstituted DRa and DRb and post-PCR products at 4°C (*see* **Note 1**).

2.3. Hybridization

1. Spoligotyping membrane (Ocimum Biosolutions Inc., formerly Isogen Biosciences B.V., Hyderabad, India).
2. MN45 miniblotter and support cushions (Immunetics, Inc., Boston, MA).
3. Rotating hybridization oven.

2.4. Detection

1. 500 U streptavidin-peroxidase conjugate (Roche Diagnostics, Indianapolis, IN), resuspended in 1 mL H_2O.
2. Enhanced chemiluminescence (ECL) detection reagents 1 and 2 (GE Healthcare Life Sciences, Piscataway, NJ).
3. X-ray film.
4. X-ray film developer.

3. Methods

The spoligotyping assay is currently performed by one of two methods. The most commonly employed method (**Fig. 1**) utilizes a nylon membrane to which 43 different oligonucleotides, corresponding to the 43 unique spacer sequences, have been individually bound *(3)*. A second method utilizes a high-throughput, multianalyte flow system (Luminex) *(22)*, permitting analysis of high numbers of strains without the need for membranes.

Detailed instructions on how to manufacture spoligotyping membranes have been previously published *(23)*. Commercially prepared spoligotyping membranes are commonly used (*see* **Note 2**). A wide variety of samples is suitable for the PCR reaction. Extracted DNA, heat-killed cell suspensions from growth medium, and even primary specimens have been successfully used as templates in PCR reactions *(3)*.

3.1. PCR Amplification of the DR Region

1. Prepare a 25-µL PCR reaction using 1–5 µL of cell suspension or 0.5–1 µL of extracted genomic DNA. A wide range of template concentrations seem suitable for DR region amplification.
2. Use the following PCR conditions: 3 min at 96°C, followed by 20 (extracted DNA) to 30 (cell suspension) cycles of 1 min at 96°C, 1 min at 55°C, and 30 s at 72°C, final extension of 5 min at 72°C (*see* **Note 3**).

3.2. Hybridization of PCR Samples to Spoligotyping Membrane

1. Add 150 µL of 2X SSPE/0.1% SDS to each tube containing the 20–25 µL post-PCR products (*see* **Note 4**).
2. Heat denature the diluted PCR products for 10 min at 100°C and cool on ice water for 2 min.
3. Wash the spoligotyping membrane for 5 min at 60°C in 2X SSPE/0.1% SDS.
4. Place the membrane and a support cushion into the miniblotter in such a way that the slots are oriented perpendicular to the line pattern of the applied oligonucleotides (*see* **Note 5**).
5. Fill the slots of the diluted PCR product (avoid air bubbles) and hybridize for 1 h at 60°C (*see* **Note 6**).

3.3. Posthybridization Steps

1. Following hybridization, remove the samples from the miniblotter by aspiration.
2. Wash the membrane twice in 2X SSPE/0.5% SDS for 10 min per wash at 60°C.
3. Place the membrane in a rolling bottle and allow it to cool to prevent inactivation of the peroxidase in the next step.

4. Add 2.5 µL of 500 U/mL streptavidin-peroxidase to 10 mL of 2X SSPE/0.5% SDS and add to roller bottle (*see* **Note 7**). Incubate the membrane in this solution for 45–60 min at 42°C with rotation (*see* **Note 8**).
5. Following this incubation, wash the membrane twice in 2X SSPE/0.5% SDS for 10 min per wash at 42°C.
6. Rinse the membrane twice with 2X SSPE for 5 min per wash at room temperature.

3.4. Chemiluminescent Detection

1. For chemiluminescent detection of hybridizing DNA, incubate the membrane for 1 min in 10 mL ECL detection reagent 1 mixed with 10 mL ECL detection reagent 2 at room temperature.
2. Briefly blot off excess ECL liquid, cover the membrane with plastic wrap, and expose to X-ray film for 2 min or longer.

3.5. Regeneration of the Membrane

The hybridized PCR product is dissociated from the membrane to regenerate the membrane for the next hybridization (*see* **Note 9**). A membrane can typically be regenerated for reuse at least 20 times.

1. Wash the membrane twice in 1% SDS at 80°C for 30 min.
2. Wash the membrane in 20 m*M* EDTA at room temperature for 15 min.
3. Store the membrane at 4°C sealed in a plastic bag containing 10 mL of 20 m*M* EDTA.

3.6. Data Analysis

3.6.1. Octal Code Nomenclature

The spoligotype patterns from X-ray film can either be read manually or scanned into a software package. For manual reads, it is recommended to have two people independently score the results for maximum accuracy. **Fig. 2** illustrates the process of assigning a 15-digit octal code *(5)* to a spoligotype result based on the pattern of hybridization. Spacers are grouped into triplets except for spacer 43. There is a number designation for each of the eight possible hybridization combinations for a group of three spacers as shown. The 15th digit of the octal code is either a 1 or a 0 based on the hybridization of spacer 43 alone.

3.6.2. Spoligotype Family Assignments

Once the octal code for a strain has been determined, the strain can then be assigned to a global strain family. Visual rules established as part of SpolDB *(6)* and an online software tool, SpotClust (http://www.rpi.edu/~bennek/EpiResearch) *(15)*, are available for aiding in the assignment of a spoligotype to one of the global strain families. The family assignment is useful for producing an overall picture of the strain diversity in a given population, for tracking changes in the TB population over time, and for comparing TB diversity between populations or areas. The criteria used to define spoligotype global families have been validated through comparison with MIRU data *(24)*.

The online version of the most recent international spoligotype database, SpolDB4, is called SIT VIT (http://www.pasteur-guadeloupe.fr:8081/SITVITDemo/). The user can enter a spoligotype octal or binary code to search whether that spoligotype has been previously reported. If it has, a shared type (ST) number is returned. That number can then be used to produce a map showing laboratories that have previously submitted that pattern. This information has the potential to be useful in deciphering the global origin or spread of a particular spoligotype.

3.7. Expansion of the Spoligotype Assay: Examination of Additional Spacer Sequences

Since the initial spoligotyping assay was based on a miniblotter with 43 usable sample chambers, the assay was limited to 43 different spacer sequences. Researchers have explored the potential of additional spacer sequences for "expanded" or "extended" spoligotyping *(25, 26)*. The hope was that screening for additional spacer sequences would improve the differentiability of the commonly observed spoligotypes. These expanded assays have not been standardized and are not commercially available, and they have unfortunately shown limited success in improving differentiation within the commonly observed patterns.

4. Notes

1. Never store biotinylated primers or biotinylated PCR products below 4°C.
2. Validate the proper manufacture of a new spoligotyping membrane on the first use by including a series of previously characterized strains.
3. To confirm PCR amplification of DR region, run a 5-µL aliquot from the reaction on a 1% miniagarose gel. A successful PCR reaction should appear as a ladder or smear of faint bands. If no PCR reactions can be observed, check oligonuleotide stocks for degradation/incorrect concentration.
4. To minimize handling of PCR products, use a 25-µL total volume PCR reaction in a 0.5-mL tube. The 150 µL of hybridization buffer may be directly added to the tube following amplification.
5. Do not reuse plastic support cushions in miniblotter.
6. Leakage into adjoining wells usually results from a dry membrane "wicking" sample into the adjoining well. Improper placement of the support cushion or membrane can also lead to this problem. Avoid wrinkling the membrane in the miniblotter. Ensure that the miniblotter is evenly tightened. Do not completely fill or overfill wells. Hybridization fluid may transfer to adjacent wells.

7. Discard stocks of strepavidin alkaline phosphatase 6 mo after rehydration.
8. Check hybridization temperature and the temperatures of the posthybridization wash buffers. Lowering the hybridization temperature and stringent washes from 60 to 55°C may help and does not add to any background problems or nonspecific hybridization.
9. With proper handling and storage, spoligotyping membranes can be reused 30 or more times.

References

1. Clark, C. M., Driver, C. R., Munsiff, S. S., Driscoll, J. R., Kreiswirth, B. N., Zhao, B., et al. (2006). Implementing universal TB genotyping in a tuberculosis control program, New York City, 2001–2003. *Emerg. Infect. Dis.* **12**, 719–724.
2. van Embden, J. D. A., Cave, M. D., Crawford, J. T., Dale, J. W., Eisenach, K. D., Gicquel, B., et al. (1993). Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *J. Clin. Microbiol.* **31**, 406–409.
3. Kamerbeek, J., Schouls, L., Kolk, A., van Agterveld, M., van Soolingen, D., Kuijper, S., et al. (1997). Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J. Clin. Microbiol.* **35**, 907–914.
4. van Embden, J. D. A., Gorkom, T., Kremer, K., Jansen, R., van der Zeijst, B. A. M., and Schouls, L. M. (2000). Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria. *J. Bacteriol.* **182**, 2393–2401.
5. Dale, J. W., Brittain, D., Cataldi, A. A., Cousins, D., Crawford, J. T., Driscoll, J., et al. (2001). Spacer oligonucleotide typing of bacteria of the *Mycobacterium tuberculosis* complex: recommendations for standardised nomenclature. *Int. J. Tuberc. Lung Dis.* **5**, 216–219.
6. Brudey, K., Driscoll, J. R., Rigouts, L., Prodinger, W. M., Gori, A., Al-Hajoj, S. A., et al. (2006). An appraisal of the geographic prevalence of major genotyping families of *Mycobacterium tuberculosis* complex through the updated SpolDB4 database. *BMC Microbiol.* **6**, 23.
7. Glynn, J. R., Whiteley, J., Bifani, P. J., Kremer, K., and van Soolingen, D. (2002). Worldwide occurrence of Beijing/W strains of *Mycobacterium tuberculosis*: a systematic review. *Emerg. Infect. Dis.* **8**, 843–849.
8. de Viedma, D. G., Rodriguez, N. A., Andres, S., Lirola, M. M., Ruiz-Serrano, M. J., and Bouza, E. (2006). Evaluation of alternatives to RFLP for the analysis of clustered cases of tuberculosis. *Int. J. Tuberc. Lung Dis.* **10**, 454–459.
9. Gopaul, K. K., Brown, T. J., Gibson, A. L., Yeates, M. D., and Drobniewski, F. A. (2006). Progression toward an improved DNA amplification-based typing technique in the study of *Mycobacterium tuberculosis* epidemiology. *J. Clin. Microbiol.* **44**, 2492–2498.
10. Streicher, E. M., Victor, T. C., van der Spuy, G., Sola, C., Rastogi, N., van Helden, P. D., et al. (2007). Spoligotype signatures in the *Mycobacterium tuberculosis* complex. *J. Clin. Microbiol.* **45**, 237–240.
11. Smith, N. H., Kremer, K., Inwald, J., Dale, J., Driscoll, J. R., Gordon, S. V., et al. (2006). Ecotypes of the *Mycobacterium tuberculosis* complex. *J. Theor. Biol.* **239**, 220–225.
12. Gori, A., Esposti, A. D., Bandera, A., Mezzetti, M., Sola, C., Marchetti, G., et al. (2005). Comparison between spoligotyping and IS*6110* fragment length polymorphisms in molecular genotyping analysis of *Mycobacterium tuberculosis* strains. *Mol. Cell. Probes* **19**, 236–244.
13. van Soolingen, D., Qian, L., de Haas, P. E. W., Douglas, J. T., Traore, H., Portaels, F., et al. (1995). Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of East Asia. *J. Clin. Microbiol.* **33**, 3234–3238.
14. Scott, A. N., Menzies, D., Tannenbaum, T., Thibert, L., Kozak, R., Joseph, L., et al. (2005). Sensitivities and specificities of spoligotyping and mycobacterial interspersed repetitive unit-variable number tandem repeat typing methods for studying molecular epidemiology of tuberculosis. *J. Clin. Microbiol.* **43**, 89–94.
15. Vitol, I., Driscoll, J., Kreiswirth, B., Kurepina, N., and Bennett, K. P. (2006). Identifying *Mycobacterium tuberculosis* complex strain families using spoligotypes. *Infect. Genet. Evol.* **6**, 491–504.

16. Ellis, B. A., Crawford, J. T., Braden, C. R., McNabb, S. J., Moore, M., Kammerer, S., et al. (2002). Molecular epidemiology of tuberculosis in a sentinel surveillance population. *Emerg. Infect. Dis.* 8, 1197–1209.
17. National TB Controllers Association/CDC Advisory Group on Tuberculosis Genotyping. (2004). *Guide to the Application of Genotyping to Tuberculosis Prevention and Control.* Atlanta, GA: U.S. Department of Health and Human Services, CDC. Available at: http://www.cdc.gov/TB/genotyping/manual.htm.
18. Gori, A., Bandera, A., Marchetti, G., Degli Esposti, A., Catozzi, L., Nardi, G. P., et al. (2005). Spoligotyping and *Mycobacterium tuberculosis. Emerg. Infect. Dis.* **11**, 1242–1248.
19. Sintchenko, V., and Gilbert, G. L. (2007). Utility of genotyping of *Mycobacterium tuberculosis* in the contact investigation: a decision analysis. *Tuberculosis* **87**, 176–184.
20. Niven, B., Driscoll, J., Bifani, P., Glaser, T., and Munsiff, S. (2000). Use of spoligotype analysis to detect laboratory cross-contamination. *J. Clin. Microbiol.* **38**, 3200–2004.
21. Bifani, P., Moghazeh, S., Shopsin, B., Driscoll, J., Ravikovitch, A., and Kreiswirth, B. N. (2000). Molecular characterization of *Mycobacterium tuberculosis* H37Rv/Ra variants: distinguishing the mycobacterial laboratory strain. *J. Clin. Microbiol.* **38**, 3200–3204.
22. Cowan, L. S., Diem, L., Brake, M. C., and Crawford, J. T. (2004). Transfer of a *Mycobacterium tuberculosis* genotyping method, spoligotyping, from a reverse line-blot hybridization membrane-based assay to the Luminex multianalyte profiling system. *J. Clin. Microbol.* 42, 474–7.
23. Molhuizen, H. O. F., Bunschoten, A. E., Schouls, L. M., and van Embden, J. D. A. (1998). Rapid detection and simultaneous strain differentiation of *Mycobacterium tuberculosis* complex bacteria by spoligotyping, in *Mycobacteria Protocols* (Parish, T., and Stroker, N. G, eds.), Humana, Totowa, NJ, pp. 381–394.
24. Ferdinand, S., Valetudie, G., Sola, C., and Rastogi, N. (2004). Data mining of *Mycobacterium tuberculosis* complex genotyping results using mycobacterial interspersed repetitive units validates the clonal structure of spoligotyping-defined families. *Res. Microbiol.* **155**, 647–654.
25. Javed, M. T., Aranaz, A., de Juan, L., Bezos, J., Romero, B., Alvarez, J., et al. (2007). Improvement of spoligotyping with additional spacer sequences for characterization of *Mycobacterium bovis* and *M. caprae* isolates from Spain. *Tuberculosis* **87**, 437–445.
26. van der Zanden, A. G. M., Kremer, K., Schouls, L. M., Caimi, K., Cataldi, A., Hulleman, A., et al. (2002). Improvement of differentiation and interpretability of spoligotyping for *Mycobacterium tuberculosis* complex isolates by introduction of new spacer oligonucleotides. *J. Clin. Microbiol.* **40**, 4628–4639.

Chapter 11

Multilocus Sequence Typing

Ana Belén Ibarz Pavón and Martin C.J. Maiden

Abstract

Multilocus sequence typing (MLST) was first proposed in 1998 as a typing approach that enables the unambiguous characterization of bacterial isolates in a standardized, reproducible, and portable manner using the human pathogen *Neisseria meningitidis* as the exemplar organism. Since then, the approach has been applied to a large and growing number of organisms by public health laboratories and research institutions. MLST data, shared by investigators over the world via the Internet, have been successfully exploited in applications ranging from molecular epidemiological investigations to population biology and evolutionary analyses. This chapter describes the practical steps in the development and application of an MLST scheme and some of the common tools and techniques used to obtain the maximum benefit from the data. Considerations pertinent to the implementation of high-capacity MLST projects (i.e., those involving thousands of isolates) are discussed.

Key words: High-throughput sequencing, MLST, population genetics, sequence types.

1. Introduction

Multilocus sequence typing (MLST) *(1)* combined a number of technical and conceptual developments of the last two decades of the 20th century to provide a universal, portable, and precise means of typing bacteria *(1–3)*. The approach owed much to the pioneering technique of multilocus enzyme electrophoresis (MLEE) (*see* **Chapter 2**), from which it acquired its name *(4)*. A key conceptual development was the recognition that bacteria do not necessarily have a clonal population structure *(5,6)*, leading to the realization that patterns of genetic exchange among bacteria, and therefore of descent, could only be resolved by the analysis of nucleotide sequence data from multiple locations of the

D.A. Caugant (ed.), *Molecular Epidemiology of Microorganisms*, Methods in Molecular Biology, Vol. 551
DOI: 10.1007/978-1-60327-999-4_11,

chromosome *(7)*. Developments in high-throughput nucleotide sequence determination and analysis permitted the generation of definitive genetic data from any locus on the chromosome of multiple isolates *(8)*. An advantage of nucleotide sequence data is that they can be disseminated via the Internet, particularly the World Wide Web *(9,10)*.

The first MLST scheme developed was for the human pathogen *Neisseria meningitidis (1)*, largely as a result of the leading role that studies of this organism had played in the development of the more sophisticated appreciation of bacterial population structure *(11–14)*. It is noteworthy that the success of this scheme was, to a great extent, due to its immediate acceptance by the wide community of researchers working on pathogenic *Neisseria*. This was due to the fact that the scheme was developed and promoted by a consortium of leading researchers in the fields of meningococcal epidemiology and population biology. Cooperation and collaboration continue to be cornerstones of successful MLST schemes.

MLST has since been applied to a number of different bacteria and eukaryotic organisms as a tool for the epidemiological analysis and surveillance of pathogens as well as to investigate their population structure and evolution. MLST has also been deployed in studies of the population structure of nonpathogenic bacteria *(2)*.

MLST provides a number of advantages over other typing approaches. First, it uses sequence data and can therefore detect changes at the DNA level that are not apparent by phenotypic approaches, such as serotyping, and by MLEE that uses the migration rate of proteins in starch gels. Second, it is a generic technique that can be readily reproduced and does not require access to specialized reagents or training. Third, modern methods of direct nucleotide sequencing, based on the polymerase chain reaction (PCR), do not require direct access to live bacterial isolates or high-quality genomic DNA. These techniques can be performed on killed cell suspensions, avoiding all the difficulties associated with the transport and manipulation of pathogens, or on clinical samples, such as the cerebrospinal fluid or blood of a patient undergoing antibiotic therapy, from which a live bacterial isolate might be difficult to obtain. Fourth, the data generated are fully portable among laboratories and can be shared throughout the world via the Internet. Finally, the Internet can also be used to disseminate MLST methods, providing standardization of approaches *(2)*.

This chapter describes the principles behind the development and application of an MLST scheme using the methods deployed in the *Neisseria* scheme as an example. In particular, the upscaling of MLST to enable the cost-effective typing of many hundreds or thousands of isolates is discussed. The general principles are applicable to essentially all bacteria, although the utility depends

on the diversity of the population under investigation and the question asked. The chapter concludes with an overview of some of the approaches available for the basic analysis of MLST data.

2. Materials

2.1. Isolate Collection

A representative sample of the population for which the scheme is to be developed (*see* **Note 1**).

2.2. Preparation of Killed Cell Suspensions

1. Freshly grown plates of bacterial cultures.
2. Boiling water bath.
3. Sterile phosphate-buffered saline (PBS).
4. 1.5- to 2.0-mL screw-capped microcentrifuge (Eppendorf) tubes (not double-walled or skirted tubes).
5. Sterile swabs/loops.

2.3. PCR Amplification of Gene Fragments

1. DNA template.
2. Forward and reverse primers.
3. DNA polymerase enzyme (*Taq* polymerase).
4. Deoxyribonucleoside 5′-triphosphates (dNTPs).
5. Buffer solution (supplied with the enzyme).
6. Magnesium chloride (supplied with the enzyme).
7. Microtiter plates resistant to high temperatures or 0.6-mL capacity microfuge tubes.
8. Thermocycler.

2.4. Gel electrophoresis

1. Agarose.
2. Ethidium bromide: 10 mg/mL stock solution.
3. Loading buffer.
4. TBE buffer: A 10X stock (0.89*M* Tris-HCl, 0.89*M* boric acid, 20 m*M* ethylenediaminetetraacetic acid [EDTA], pH 8.3).
5. Power source.
6. UV transilluminator.

2.5. PCR Product Purification

1. 1.5-mL microcentrifuge tubes.
2. Polyethylene glycol (PEG) 8000.
3. Sodium chloride.
4. Ethanol 70%.
5. Benchtop centrifuge.

2.6. Sequencing Reactions

1. Purified PCR products (DNA template).
2. Forward and reverse primers.
3. Sequencing kit containing DNA polymerase and labeled dNTPs.
4. Microtiter plates or 0.6-µL tubes.
5. Thermocycler.
6. DNA sequencer.

2.7. Purification of Sequencing Products

1. 1.5-mL microcentrifuge tubes.
2. 3*M* sodium acetate, pH 4.6.
3. Ethanol, 95% and 70%.
4. Benchtop centrifuge.

3. Methods

3.1. Killed Cell Suspensions

1. Heat the water bath until it boils.
2. Clearly label the screw-capped microcentrifuge tubes. Ensure that these labels will not come off during the heating step.
3. Dispense 0.5 mL of PBS in each microcentrifuge tube.
4. Make very thick suspensions of organisms by sweeping colonies from each culture plate using a swab or a loop and emulsifying in the PBS in the tubes.
5. Place the tubes in the boiling water bath and leave for 20 min.
6. Store the samples at –20°C. These samples are, in principle, killed and stable at room temperature. Once lack of viability has been confirmed, they can be handled in the laboratory and distributed as noninfectious material (*see* **Note 2**).

3.2. PCR Amplification (see Note 3)

1. Initialization: The reaction mix is heated to 94°C for 1 min to denature the DNA.
2. The following steps are repeated for 25–30 cycles:
   a. Denaturation at 94°C for 30 s.
   b. Primer annealing at 50–60°C for 30 s. This allows the primers to bind to the template DNA.
   c. Extension at 72°C. During this step, the *Taq* polymerase uses the dNTPs to synthesize a new DNA strand complementary to the template. The duration of this step depends on the length of the fragment that is to be amplified.

3. Final elongation at 72°C for 5–10 min to ensure that all the fragments are fully extended.
4. The reaction should be held at 4°C until removed from the thermocycler.

3.3. Agarose Gel Electrophoresis (see Note 4 and ref. 15)

1. Prepare a 1% (w/v) agarose gel by adding 1 g of agarose to 100 mL of TBE buffer.
2. Heat in a microwave until boiling.
3. Leave it to cool for 2–3 min.
4. Add 5 μL of ethidium bromide.
5. Insert the gel comb and wait until is solid.
6. Fill in the electrophoresis tank with TBE.
7. Insert the gel into the tank and remove the combs.
8. Mix 5 μL of PCR product with 2 μL of loading buffer.
9. Connect the gel tank to the power source.
10. Set the voltage to 140 V and leave it running for 15–20 min.
11. Visualize the gel using a UV transilluminator.

3.4. Purification of Amplicons (see Note 5)

1. Transfer the contents of each PCR tube into labeled 1.5-mL Eppendorf tubes. If microtiter plates are used, this step can be omitted.
2. Add 60 μL of 20% (w/v) PEG 8000, 2.5*M* sodium chloride to each tube and mix. Incubate for 30 min at room temperature.
3. Pellet the PCR products by spinning in a centrifuge at maximum speed for 15 min. For microtiter plates, spin for 1 h at 2,750*g*.
4. Discard the supernatant and wash the DNA pellet by adding 0.5 mL of 70% ethanol and spin at maximum speed for a further 5 min. For microtiter plates, add 150 μL of 70% ethanol and spin for 10 min at 2,750*g*. Repeat this step twice when using plates.
5. Discard the supernatant and dry pellets in the vacuum dryer. Microtiter plates can be dried by spinning upside down for 1 min at 500*g*.

3.5. Nucleotide Sequence Extension Reactions (see Note 6)

1. Mix the primer, template, and sequencing reagents in the optimized proportions.
2. Perform the extension reactions in a thermocycler, first conducting denaturation at 96°C for 1 min.

3. The following steps are repeated for 25 cycles: 96°C for 10 s, 50°C for 5 s, 60°C for 40 min.
4. Maintain the reaction at 4°C until removed from the thermocycler.

3.6. Purification of Sequencing Products

1. Transfer the contents of each PCR tube into labeled 1.5-mL Eppendorf tubes. If microtiter plates are used, this step can be omitted.
2. Add 2 µL of 3*M* sodium acetate and 50 µL of 95% ethanol to each tube and mix. Incubate for 45 min at room temperature.
3. Pellet the PCR products by spinning in a centrifuge at maximum speed for 15 min. For microtiter plates, spin for 1 h at 2,750*g*.
4. Discard the supernatant and wash the DNA pellet by adding 0.5 mL of 70% ethanol and spin at maximum speed for a further 5 min. For microtiter plates, add 150 µL of 70% ethanol and spin for 10 min at 2,750*g*.
5. Discard the supernatant and dry pellets in the vacuum dryer. Microtiter plates can be dried by spinning upside down for 1 min at 500*g*.
6. For separation and detection of extension products, *see* **Note 7**.

3.7. Data Management

3.7.1. Data Assembly

A variety of commercial and open source software packages are available for the assembly and editing of sequence chromatograms into compiled edited sequences, including the well-known Staden and GCG packages *(16,17)*. Specialist software for the compilation and analysis of MLST data is also available, for example, the START software package *(18)*. These packages allow many hundreds or even thousands of samples to be processed cost-effectively and rapidly. Inexpensive Linux-based software *(19)*, as well as commercial solutions, are available. The use of Internet-based databases and analytical tools designed for MLST analysis can automatically designate sequence type (ST) and clonal complex as well as facilitate storage and access to the data via Internet. This procedure is described in detail in **Chapter 21**.

The sequence type analysis and retrieval system (STARS) is specifically designed for the assembly of MLST data (http://www.cbrg.ox.ac.uk/~mchan/stars/). It uses PREGAP4 and GAP4 from the Staden package *(16)* to automatically assemble a large number of sequences, which can be retrieved and edited. For known alleles and STs, designation can be done directly from the STARS interface by interrogating an MLST database.

3.7.2. Data Storage

The maintenance of curated, Web-accessible databases is a key feature of MLST schemes. These databases act as dictionaries that allow bacterial isolates to be compared worldwide *(2)*. Database management is therefore central to the endeavor. The key part

of MLST databases is comprised of the allele sequences linked to MLST allele numbers for each locus and the definition of STs. In some cases, it may be appropriate to include information on higher-order organization of STs into clonal complexes or lineages in this database as is done with the *Neisseria* MLST Allelic Profile/ST Database. These data can then be linked to isolate databases that contain isolate specific information. It is important that there is a separation between the databases containing the allele and ST data and isolate data as many isolates will contain the same alleles or STs *(9)*.

3.8. Data Analysis

3.8.1. Analysis of MLST Data

The first question to be addressed with an MLST data set is whether the data conform to the clonal model of population structure. Clonal population structure is an inevitable consequence of asexual reproduction combined with diversity reduction events, such as periodic selection and sequential bottlenecks *(20)*. If an organism is clonal, then the analysis is greatly simplified as conventional phylogenetic trees can be employed. Clonality can be investigated by the congruence test *(21)*, which is based on the observation that, in a clonal population, the phylogenetic signal observed at different loci is the same or congruent *(22)*.

Most bacteria that have been analyzed by MLST are, however, nonclonal by the congruence test. For such organisms the clonal complex is a useful concept that groups genetically related organisms. Clonal complexes comprise groups of related STs that are likely to derive from a common ancestor. Currently, the designation of clonal complex is pragmatic and to an extent varies with different bacteria, but the important issue is that the grouping is consistent with what is known and understood about the genealogy of the organism. The BURST (based upon related sequences) algorithm is a rapid and effective algorithm that can be used to assign the central genotype of clonal complexes. The eBURST program *(23)* groups STs into groups according to user-defined criteria of a number of alleles in common to at least one other member of the group. The central genotype of a BURST group will be the one with the highest number of single-locus variants (SLVs). This will often coincide with the one most frequently isolated and therefore gives some biological meaning to the future designation of the clonal complex. The eBURST program and instructions can be found at http://eburst.mlst.net/. A number of clustering algorithms, such as the unweighted pair group method with arithmetic mean (UPGMA) *(24)* or split decomposition *(25)* can be used to cluster STs and reinforce the results obtained using eBURST.

3.8.2. Applying the Clone Complex Model

It is possible to rationalize the clonal complex structure of many bacteria in terms of the "epidemic clone model" *(5)* of bacterial population structure or modifications of it. Within this frame-

work, high prevalence of a single ST indicates the presence of a fast-spreading new clone from which variants are developing. In the absence of a formal means of defining such clones, it is necessary to implement a rational definition that will command support from the scientific community analyzing these bacteria. It is advisable to designate a committee of experts who ultimately decide on the management and nomenclature issues raised by the scheme.

3.9. High-Throughput MLST

One of the great advantages of MLST is its scalability from a single bacterial isolate to many hundreds or even thousands of samples. Upscaling of MLST is essential for large-scale studies and brings with it appreciable advantages in terms of reducing costs. Automation reduces staff input, and bulk purchase of reagents brings substantial cost savings. While automation brings substantial benefits, it does require substantial commitment and investment. During the setup process the various sections of the data production and analysis pipeline have to be analyzed and kept under review; potential bottlenecks can then be identified and handled. Any process is only as efficient and rapid as its least-efficient and slowest step. PCR and sequencing reactions can be automated by investing in a robotic platform that saves personnel time and minimizes error *(26,27)* or at least ensures that any error is deterministic rather than stochastic. A number of fast and reliable methods exist for the purification of amplification products that can be incorporated into the robotic platform, although consumables for these types of systems are often expensive. The PEG precipitation for PCR products and sodium acetate/ethanol precipitation for sequence reactions are highly cost-effective, but are time consuming and require investment in centrifugation equipment capable of sedimenting material in microtiter plates.

Optimization of the sequence reactions and the use of a centralized sequencing facility can further reduce costs as the use of reagents can be minimized, and costs can be further reduced by bulk purchase *(28)*. If automation is to be used, it is important to recognize that the processes are more akin to those found in industrial rather than conventional biological research organizations. Robotic equipment works most effectively when it is regularly used to perform highly repetitive operations. Once the equipment is working on a given application, a process that often requires appreciable investment of time and effort, temptation to further improve operation by minor modification should be resisted. Such attempts prevent the exploitation of the equipment efficiently and are at least as likely to degrade as to enhance the performance of the equipment.

3.10. Applications of MLST Data

3.10.1. Application to Public Health

Public health laboratories use MLST routinely for the characterization of clinical specimens *(29,30)*. For the meningococcus, for example, the information obtained has proven to be invaluable for the understanding and management of disease outbreaks *(31,32)*, epidemiological surveillance *(33,34)*, and the monitoring of public health interventions. Its application to clinical specimens has obvious implications for diagnosis and clinical management of cases caused by an organism that is notoriously difficult to isolate microbiologically from patients undergoing antibiotic therapy *(35–37)*.

3.10.2. Evolutionary and Population Genetic Analyses

MLST data have been used in a wide variety of applications, including evolutionary and population analysis of bacterial species, but to date they have been mostly used in molecular epidemiological studies of bacterial pathogens. Molecular epidemiology employs genetic techniques to characterize isolates of infectious agents or identify their presence and characteristics from clinical specimens. By this means their distribution and spread can be monitored, and if necessary, health interventions can be implemented. MLST has been applied to many bacteria, as recently reviewed *(2)*. MLST data can also be used to investigate the population structure of bacterial populations at different levels (e.g., temporal stratification or geographic distribution) as this can help to understand the transmission route of the infectious agent *(38)*. For this purpose, the analysis of molecular variance (AMOVA) *(39)* can be used to calculate the F statistic (F_{ST}) *(40)*, which measures the amount of genetic exchange that takes place among different groups of organisms. The Mantel test can be used to investigate the correlation between genetic and geographic distance, that is, whether isolates obtained from geographically close locations are more closely related to those found on more distant geographic areas *(38)*. Both tests can be easily implemented using the Arlequin software package *(39)*, which can be downloaded from http://lgb.unige.ch/arlequin/.

4. Notes

1. The isolates examined must be carefully chosen with a number of criteria in mind: They should represent the known genetic diversity of the population analyzed (which itself should be carefully defined); they should represent a variety of sources or environments from which the organism is often isolated; and they should be collected from a variety of geographic locations and an appropriate time frame.

2. The crucial step in this method is the rapid inactivation of cellular nucleases once the cells have been lysed.
3. In an MLST scheme, PCR conditions are ideally the same for all loci. This should be straightforward if primers are designed to have similar melting temperatures T_m and if the DNA fragments to be amplified are of similar lengths. Optimization is likely to be needed for novel primers.
4. Standard agarose gel electrophoresis can be employed to check that the amplification reactions have been successful and that amplicons of the expected size have been produced. It is recommended to check all the samples during the optimization period, but when the MLST scheme is fully developed and routinely applied on a large scale, only occasional verification is necessary *(15)*.
5. A variety of methods for purification are available, including many commercial kits. However, the purification method described here is an effective and inexpensive noncommercial method based on sodium chloride and PEG differential purification.
6. It is an absolute requirement for accurate sequence determination that sequence information from both DNA strands is used to compile the final sequence, so PCR reactions "forward" and "reverse" are required for each DNA molecule to be sequenced. The reactions are easily performed with proprietary kits that contain all of the necessary components, requiring only template DNA and specific primer to be added. Some local optimization is likely to be required for the primers and reagents used.
7. A variety of commercial instruments is available for the separation of extension reaction products, and a description of their operation is beyond the scope of this chapter. In most cases they are capillary based and generally operated by central sequencing facilities as they are high-cost assets that, to be cost-effective, have to be used on very large numbers of samples, usually representing a wide variety of applications. Although smaller-scale instruments suitable for the use by single laboratories are available, they are usually much more expensive to run. Commercial companies also offer sequencing services.

References

1. Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U S A* **95**, 3140–3145.
2. Maiden, M. C. (2006). Multilocus sequence typing of bacteria. *Annu. Rev. Microbiol.* **60**, 561–588.
3. Urwin, R., and Maiden, M. C. (2003). Multilocus sequence typing: a tool for global epidemiology. *Trends Microbiol.* **11**, 479–487.
4. Selander, R. K., Caugant, D. A., Ochman, H., Musser, J. M., Gilmour, M. N., and Whittam, T. S. (1986). Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Appl. Environ. Microbiol.* **51**, 837–884.

5. Maynard Smith, J., Smith, N. H., O'Rourke, M., and Spratt, B. G. (1993). How clonal are bacteria?. *Proc. Natl. Acad. Sci. U S A* **90**, 4384–4388.
6. Maynard Smith, J., Dowson, C. G., and Spratt, B. G. (1991). Localized sex in bacteria. *Nature* **349**, 29–31.
7. Maynard Smith, J., Feil, E. J., and Smith, N. H. (2000). Population structure and evolutionary dynamics of pathogenic bacteria. *BioEssay* **22**, 1115–1122.
8. Maiden, M. C. J. (2000). High-throughput sequencing in the population analysis of bacterial pathogens. *Int. J. Med. Microbiol* **290**, 183–190.
9. Jolley, K. A., Chan, M. S., and Maiden, M. C. (2004). mlstdbNet—distributed multi-locus sequence typing (MLST) databases. *BMC Bioinformatics* **5**, 86.
10. Chan, M. S., Maiden, M. C., and Spratt, B. G. (2001). Database-driven multi locus sequence typing (MLST) of bacterial pathogens. *Bioinformatics.* **17**, 1077–1083.
11. Caugant, D. A., Frøholm, L. O., Bovre, K., Holten, E., Frasch, C. E., Mocca, L. F., et al. (1986). Intercontinental spread of a genetically distinctive complex of clones of *Neisseria meningitidis* causing epidemic disease. *Proc. Natl. Acad. Sci U S A* **83**, 4927–4931.
12. Olyhoek, T., Crowe, B. A., and Achtman, M. (1987). Clonal population structure of *Neisseria meningitidis* serogroup A isolated from epidemics and pandemics between 1915 and 1983. *Rev. Infect. Dis.* **9**, 665–682.
13. Zhang, Q. Y., Jones, D. M., Saez Nieto, J. A., Perez Trallero, E., and Spratt, B. G. (1990). Genetic diversity of penicillin-binding protein 2 genes of penicillin-resistant strains of *Neisseria meningitidis* revealed by fingerprinting of amplified DNA. *Antimicrob. Agents Chemother.* **34**, 1523–1528.
14. Feavers, I. M., Heath, A. B., Bygraves, J. A., and Maiden, M. C. (1992). Role of horizontal genetic exchange in the antigenic variation of the class 1 outer membrane protein of *Neisseria meningitidis. Mol. Microbiol.* **6**, 489–495.
15. Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982). Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
16. Staden, R. (1996). The Staden sequence analysis package. *Mol. Biotechnol.* **5**, 233–241.
17. Womble, D. D. (2000). GCG: The Wisconsin package of sequence analysis programs. *Methods Mol. Biol.* **132**, 3–22.
18. Jolley, K. A., Feil, E. J., Chan, M. S., and Maiden, M. C. (2001). Sequence type analysis and recombinational tests (START). *Bioinformatics.* **17**, 1230–1231.
19. Field, D., Tiwari, B., and Snape, J. (2005). Bioinformatics and data management support for environmental genomics. *PLoS Biol.* **3**, e297.
20. Gupta, S., and Maiden, M. C. J. (2001). Exploring the evolution of diversity in pathogen populations. *Trends Microbiol.* **9**, 181–192.
21. Holmes, E. C., Urwin, R., and Maiden, M. C. J. (1999). The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis. Mol. Biol. Evol.* **16**, 741–749.
22. Dykhuizen, D. E., Polin, D. S., Dunn, J. J., Wilske, B., Preac Mursic, V., Dattwyler, R. J., et al. (1993). *Borrelia burgdorferi* is clonal: implications for taxonomy and vaccine development. *Proc. Natl. Acad. Sci U S A* **90**, 10163–10167.
23. Feil, E. J., Li, B. C., Aanensen, D. M., Hanage, W. P., and Spratt, B. G. (2004). eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J. Bacteriol.* **186**, 1518–1530.
24. Kumar, S., Tamura, K., and Nei, M. (2004). MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinform.* **5**, 150–163.
25. Huson, D. H. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics.* **14**, 68–73.
26. Jefferies, J., Clarke, S., Diggle, M., Smith, A., Dowson, C., and Mitchell, T. (2003). Automated pneumococcal MLST using liquid-handling robotics and a capillary DNA sequencer. *Mol. Biotechnol.* **24**, 303–308.
27. Clarke, S. C. (2002). Nucleotide sequence-based typing of bacteria and the impact of automation. *Bioessays* **24**, 858–862.
28. Diggle, M. A., and Clarke, S. C. (2002). What a load of old sequence!!!. *J. Clin. Microbiol.* **40**, 2707.
29. Kriz, P., Kalmusova, J., and Felsberg, J. (2002). Multilocus sequence typing of *Neisseria meningitidis* directly from cerebrospinal fluid. *Epidemiol. Infect.* **128**, 157–160.
30. Diggle, M. A., Bell, C. M., and Clarke, S. C. (2003). Nucleotide sequence-based typing of meningococci directly from clinical samples. *J. Med. Microbiol.* **52**, 505–508.
31. Bygraves, J. A., Urwin, R., Fox, A. J., Gray, S. J., Russell, J. E., Feavers, I. M., et al. (1999). Population genetic and evolutionary approaches to the analysis of *Neisseria meningitidis* isolates belonging to the ET-5 complex. *J. Bacteriol.* **181**, 5551–5556.
32. Feavers, I. M., Gray, S. J., Urwin, R., Russell, J. E., Bygraves, J. A., Kaczmarski, E. B., et al. (1999). Multilocus sequence typing and antigen gene sequencing in the investigation

of a meningococcal disease outbreak. *J. Clin. Microbiol.* **37**, 3883–3887.

33. Brehony, C., Jolley, K. A., and Maiden, M. C. (2007). Multilocus sequence typing for global surveillance of meningococcal disease. *FEMS Microbiol. Rev.* **31**, 15–26.
34. Zhang, X. B., Shao, Z. J., Yang, E., Xu, L., Xu, X. Y., Li, M. C., et al. (2007). Molecular characterization of serogroup C *Neisseria meningitidis* isolated in China. *J. Med. Microbiol.* **56**, 1224–1229.
35. Cartwright, K., Reilly, S., White, D., Stuart, J., Begg, N., and Constantine, C. (1993). Management of early meningococcal disease. *Lancet* **342**, 985–986.
36. Cartwright, K., Reilly, S., White, D., and Stuart, J. (1992). Early treatment with parenteral penicillin in meningococcal disease. *BMJ.* **305**, 143–147.
37. Ni, H., Knight, A. I., Cartwright, K., Palmer, W. H., and McFadden, J. (1992). Polymerase chain reaction for diagnosis of meningococcal meningitis. *Lancet* **340**, 1432–1434.
38. Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27**, 209–220.
39. Schneider, S., Roessli, D., and Excoffier, L. (2000). Arlequin Version 2.000: A Software for Population Genetic Data Analysis. University of Geneva, Geneva, Switzerland.
40. Wright, S. (1951). The genetical structure of populations. *Ann. Eugen.* **15**, 323–354.

Chapter 12

Multiple Locus Variable Number of Tandem Repeats Analysis

Gilles Vergnaud and Christine Pourcel

Abstract

Genotyping of bacteria through typing of loci containing a variable number of tandem repeats (VNTR) might become the gold standard for many pathogens. The development of genome sequencing has shown that such sequences were present in every species analyzed, and that polymorphism exists in at least a fraction of them. The length of these repetitions can vary from a single nucleotide to a few hundreds. This has implications for both the techniques used to measure the repeat number and the level of variability. In addition, tandem repeats can be part of coding regions or be intergenic and may play a direct role in the adaptation to the environment, thus having different observed evolution rates. For these reasons the choice of VNTR when setting a multiple-loci VNTR analysis (MLVA) assay is important. Although reasonable discrimination can be achieved with the typing of six to eight markers, in particular in species with high genomic diversity, it may be necessary to type 20 to 40 markers in monomorphic species or if an evolutionary meaningful assay is needed. Homoplasy (in the present context, two alleles containing the same repeat copy number in spite of a different history) is then compensated by the analysis of multiple markers. Finally, even if the underlying principles are relatively simple, quality standards must be implemented before this approach is widely accepted, and technology issues must be resolved to further lower the typing costs.

Key words: Database, genotyping, MLVA, MLVAbank, services on the Web, VNTR.

1. Introduction

The globalization of the world economy, with a dramatic increase of traveling of human beings and exchange of goods, is also as a consequence globalizing the spread of pathogens and associated infectious diseases. As a result, there is a growing requirement for tools enabling the real-time accountability and tracing of these

D.A. Caugant (ed.), *Molecular Epidemiology of Microorganisms*, Methods in Molecular Biology, Vol. 551
DOI: 10.1007/978-1-60327-999-4_12,

pathogens. Such tools should satisfy a number of criteria. They should be "low cost" so that any new isolate can be routinely typed. They should be not too demanding from a technological point of view, so that partial typing, or first-level low-resolution typing, can be performed in any microbiology laboratory and not only in dedicated high-throughput facilities. The resulting data should open the way to the making of large-scale databases shared across the Internet, as well as small-scale databases for local surveillance within, for instance, a hospital setting. These requirements constitute one aspect of the new discipline called forensic microbiology.

Although many methods have been developed to investigate the epidemiology of pathogens, with some described in other chapters of this book, only a few qualify as potentially of use in this context. In the present chapter, we go through multiple loci variable number of tandem repeat (VNTR) analysis (MLVA), which for a number of reasons that we illustrate is currently considered as one of the most promising technologies regarding the epidemiology of microorganisms with relatively large genomes, such as bacteria.

It is now clear that, for many reasons, including biosafety, the appropriate typing technologies will target DNA. The resulting data need to be easily storable in a digital format so that eventually worldwide coverage of the pathogen diversity can be achieved. This excludes pattern-based technologies, such as randomly amplified polymorphic DNA (*see* Chapter 4), polymerase chain reaction (PCR) amplification of multiple interspersed repeated elements (*see* Chapter 5), pulsed-field gel electrophoresis (*see* Chapter 6), amplified fragment length polymorphism (*see* Chapter 8), and insertion element (IS) typing by Southern blot hybridization (*see* Chapter 14), for instance. In such techniques, patterns produced in different laboratories can be compared only if very strict quality standards are followed. The use of polymorphic tandem repeats to discriminate biological entities is not new and is not limited to microorganisms. We do not go through the history of tandem repeat analysis for which reviews exist (e.g., ***1, 2)***. The important feature of MLVA is that the analysis of a limited number of loci provides an overview of diversity within a bacterial species.

A number of aspects specific for tandem repeats analysis must be kept in mind. First, tandem repeat loci can be very variable in terms of mutation rates, with some loci having an extreme mutation rate while others are monomorphic. At present, this behavior cannot be predicted from the sequence itself and needs to be experimentally measured by eventually typing hundreds of strains, as was done previously for human forensics-related projects. The most highly polymorphic markers that usually result from a higher rate of mutation events will have a high homoplasy level.

Such markers are sometimes called "highly informative," which is ambiguous and not necessarily correct. On the contrary, an MLVA assay that would be based solely on such markers would probably be unable to cluster strains according to their true historical proximity, as illustrated previously with *Brucella (3)*.

2. Materials

2.1. DNA Purification

1. For some bacterial species, thermolysates can be prepared in water and stored at –20°C. Long-term stability needs to be evaluated for each species. In some instances, glass beads are used to break cells.
2. Purification kit such as Qiagen DNeasy® Tissue kit. DNAs are stored at –20°C.
3. ND-1000 spectrophotometer (NanoDrop®, Labtech, France).

2.2. PCR Amplification

1. Standard *Taq* polymerase (Roche, Promega, or Invitrogen) or *Pfu* polymerase when amplifying mononucleotide repeats (*see* **Note 1**).
2. The Qiagen kit provides the "Q solution" and corresponding buffer for amplification of GC-rich DNA. Alternatively, 1*M* betain (Sigma) can be used in the PCR reaction (*see* **Note 2**).
3. Fluorescent oligonucleotides to be used with the Beckman sequencer are from Sigma-Aldrich Proligo (www.proligo.com).
4. Deoxynucleotide 5′-triphosphates (dNTPs; Eurogentec, MWG Biotech, or Amersham).
5. Different thermocyclers (including PTC 200 DNA Engine and MyCycler, Bio-Rad) have been used efficiently.
6. PCR amplifications are done in microcap tubes (rows of 8 or 12 tubes) arranged in grids accommodating up to 96 tubes in a 96-well format compatible with multichannel electronic pipeting equipment (eight-channel Biohit dispensing range 0.1 to 10 µL, 1 to 20 µL).

2.3. Agarose Gel Electrophoresis

1. Standard agarose for gels up to 3% (w/v).
2. TBE electrophoresis buffer: A 5X TBE buffer (1.1*M* Tris-HCl, 900 m*M* boric acid, 25 m*M* ethylenediaminetetraacetic acid [EDTA], pH 8.3) is prepared as a stock solution and is used at 0.5X concentration for migration. Buffer solution can be used for up to three or four runs.
3. Metaphor (FMC Bioproducts-Cambrex) is used for 4% (w/v) gel, either pure or mixed 1:1 with standard agarose.

4. 100-bp ladder or 20-bp ladder from Bio-Rad, MBI Fermentas, or Euromedex (*see* **Note 3**).
5. Electrophoresis chambers compatible with 20- to 24-cm wide gels by 20 to 40 cm long, with 40 to 50 wells, and spacing compatible with multichannel pipeting (*see* **Note 4**).
6. 10X loading buffer: 0.25% (w/v) bromophenol blue, 0.25% (w/v) xylene cyanol, 50% (v/v) glycerol, 50% (v/v) TE. TE is a 10 m*M* Tris-HCl, 1 m*M* EDTA, pH 8.0 solution.
7. Ethidium bromide: purchase as 10 mg/mL aqueous stock solution; can be stored at 4°C for years. Use at 0.25 μg/mL final concentration in 0.5X TBE buffer.

2.4. Band Size Determination

The band size can be determined using dedicated software (e.g., Quantity One v. 4.2.1 of Bio-Rad) or the BioNumerics software (Applied Maths).

3. Methods

3.1. Running an MLVA Assay

A list of bacteria for which VNTR markers have been identified and tested on collections of strains is shown in **Table 1**. In some cases these assays are very preliminary since only a small number of strains have been analyzed, and much remains to be done to definitely select an informative VNTR panel and to measure their relative value. In other cases, such as for *Mycobacterium tuberculosis*, many data are already available, and the assay can be considered more reliable. However, no consensus has yet been adopted in the corresponding scientific community (*see* **Note 5**).

If an MLVA assay is available in the literature, genotyping will consist of *(1)* preparing DNA samples; *(2)* VNTR amplification and estimation of repeat number; and *(3)* data analysis and storage.

3.1.1. The DNA Samples

Very good results have been obtained from thermolysates for *M. tuberculosis* or *Legionella pneumophila*. In some instances, MLVA typing has even been done on crude biological samples with a sufficient bacterial load *(4)*. In other cases, and for some species such as *Pseudomonas aeruginosa*, a purification step of the DNA is mandatory to get reliable PCR amplifications *(5)*.

3.1.2. PCR Amplification

1. Perform the PCR reactions in a total volume of 15 μL, containing 1–5 ng of DNA, 1X reaction buffer, 1.5 m*M* $MgCl_2$, 1 unit of *Taq* DNA polymerase, 200 μ*M* of each dNTP, 0.3 μ*M* of each flanking primer.
2. Use the following conditions for amplification: Initial denaturation cycle for 5 min at 94°C, 35 cycles of denaturation for 30 s at 94°C, annealing for 30 s at 55 to 60°C depending on

Table 1
Published MLVA Schemes

Bacteria	VNTR loci[a]	Repeat sizes (bp)	No. of isolates	Method	References
Bacillus anthracis	8	2–36	426	Sequencing gel	*(22)*
	24 (18)	9–78	32	Agarose	*(9)*
Bartonella henselae	5	45–146	44	Agarose	*(23)*
Bordetella pertussis	6	5–15	198	Sequencing gel	*(24)*
Borrelia sp.	10	2–21	41	Sequencing gel	*(25)*
Brucella sp.	8	8	22	Capillary electrophoresis	*(3)*
	18	6–134	236	Agarose	*(26)*
Burkholderia pseudomallei	32	6–15	213	Capillary electrophoresis	*(12)*
Candida albicans	3	4	100	Sequencing gel	*(27)*
Candida glabrata	6	2–3	127	Capillary electrophoresis	*(28)*
Clostridium difficile	7	3–8	86	Capillary electrophoresis	*(29)*
	7	6–17	86	Sequencing	*(30)*
Clostridium perfringens	5	6–21	112	Agarose	*(31)*
Coxiella burnetii	17	6–126	42	Agarose	*(32)*
	7	6–21		Capillary electrophoresis	*(33)*
Enterococcus faecalis	7	141–393	83	Agarose	*(34)*
Enterococcus faecium	6	121–279	392	Agarose	*(35)*
Escherichia coli O157	7	6–18	81	Sequencing	*(36)*
	7	6–30	73	Capillary electrophoresis	*(37)*
E. coli, Shigella	7	6–39	72	Capillary electrophoresis	*(38)*
Francisella tularensis	6	2–21	56	Sequencing gel	*(39)*
	25	2–23	192	Sequencing gel	*(40)*
Hemophilus influenzae	5	3–6	20	Agarose	*(41)*
Lactobacillus casei	9	6–24	63	Capillary electrophoresis	*(42)*
Legionella pneumophila	6	18–45	78	Agarose	*(43)*
	13	7–125	99	Agarose	*(16)*
Leptospira interrogans	7	34–77	51	Agarose	*(44)*
	6	36–138	39	Agarose	*(45)*
Leptospira interrogans kirschneri	5	34–77	243	Agarose	*(46)*
Listeria monocytogenes	6	9–18	25	Agarose	*(47)*
Mycobacterium avium	6	53	73	Agarose	*(48)*
	5	20–70	50	Agarose	*(49)*
Mycobacterium leprae	5	2–3	12	Sequencing	*(50)*
	9	1–27	4	Sequencing gel	*(51)*

(continued)

Table 1 (continued)

Bacteria	VNTR loci[a]	Repeat sizes (bp)	No. of isolates	Method	References
Mycobacterium tuberculosis	7	15–79	25	Agarose	*(52)*
	12 (10)	53	31	Agarose	*(53)*
	6	69	100	Agarose	*(54)*
	21 (8)	9–58	90	Agarose	*(21)*
Mycobacterium ulcerans	13	53–71	29	Agarose	*(55)*
M. ulcerans and M. marinum	7	53	66	Agilent	*(56)*
Mycoplasma mycoides	3	12–75	39	Agarose	*(57)*
Neisseria meningitidis	12	4–30	100	Capillary electrophoresis	*(58)*
	12	4–21	92	Capillary electrophoresis	*(59)*
Pseudomonas aeruginosa	7	6–115	89	Agarose	*(60)*
	15	6–129	163	Agarose/capillary electrophoresis	*(5)*
Salmonella typhimurium/ typhi	8	6–189	102	Capillary electrophoresis	*(61)*
	10 (7)	3–20	99	Agarose	*(62)*
Staphylococcus aureus	7	48–159	16	Agarose	*(63)*
Staphylococcus aureus	5	9–81	34	Agarose[b]	*(64)*
	8	9–560	200	Capillary electrophoresis[b]	*(65)*
Streptococcus pneumoniae	16	12–60	56	Agarose	*(66)*
Streptococcus uberis	7	13–208	88	Agarose	*(67)*
Salmonella enterica	10	6–232	50	Agarose	*(68)*
Shigella sonnei	26	6–168	536	Capillary electrophoresis	*(69)*
Theileria parva (protozoa)	60	2–21	20	Agar/spreadex	*(70)*
Xylella fastidiosa	7	7–9	27	Agarose	*(71)*
Yersinia pestis	25	9–60	3 + 180	Agarose	*(9,72)*
	42 (26)	1–45	24 + 156	Sequencing gel	*(73,74)*

[a]In parentheses are indicated the number of VNTR not previously published.
[b]Not a bona fide MLVA assay since the five loci used are coamplified to produce a multiband pattern, and no analysis is done on a locus-by-locus basis.

the primers, and elongation for 45 to 60 s at 72°C followed by a final elongation step for 10 min at 72°C.

3.1.3. Agarose Gel Electrophoresis

The length of PCR products can be estimated by different approaches. Importantly, the required accuracy is directly related to the repeat unit size of the loci to be used. Obviously, if the markers used for MLVA typing have a repeat unit size of more

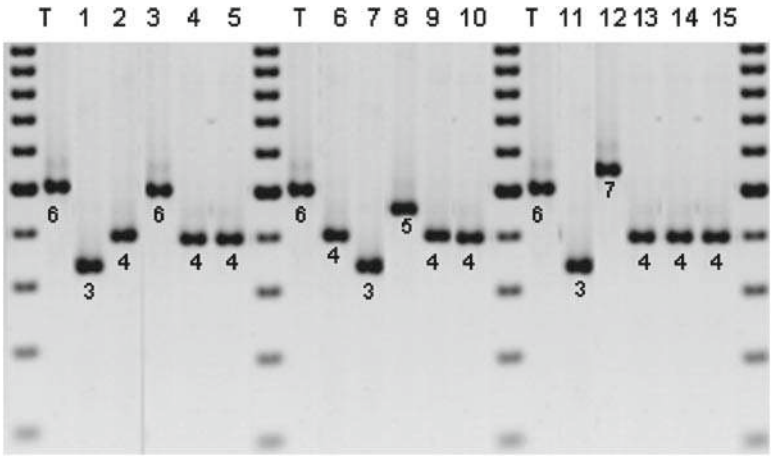


Fig. 1. Agarose gel electrophoresis of VNTR amplicons. PCR products of the *M. tuberculosis* Mtub 39 VNTR (58-bp repeat) on the reference strain H37Rv (T) and 15 isolates. The size marker is loaded every six samples.

than 50 bp (as is the case for some important pathogens such as *M. tuberculosis)*, it is not necessary to achieve a precision of plus or minus 1 bp (**Fig. 1**). In theory, a precision just below ±50% of the repeat unit size is sufficient. Accordingly, different allele-calling methods are used. The most frequently used method, at least in the assay development phase, is the ordinary agarose gel. The approximately ±2% imprecision of this approach allows the typing of loci with 6-bp repeat units if the allele size does not exceed 150–200 bp (*see* **Note 6**).

1. Add 2 µL of 10X loading buffer to the PCR products. Load 2 to 3 µL of PCR products in 2 to 3% (w/v) agarose gels made of standard agarose for repeats 9 bp and larger. (for the analysis of smaller repeats, 4% (w/v) agarose gels, comprising 2% Metaphor plus 2% regular agarose can be used) (*see* **Note 7**).
2. Perform the electrophoresis in 20- to 24-cm wide gels made in 0.5X TBE buffer, run at 8 V/cm. For each group of five isolates, include a reference strain. To ensure an adequate size assignment of the PCR products, reference size markers are run every six samples (**Fig. 1**). DNA size markers routinely used are the 100-bp ladder or 20-bp ladder (for repeats 8 bp and smaller).
3. Stain the gels after the run in 0.5–1.0 µg/mL ethidium bromide for 15 to 30 min. Then, rinse the gel with water and photograph under ultraviolet illumination (*see* **Note 8**).

3.1.4. Agarose Gel Image Analysis

The PCR product sizes can be estimated from the digital image of the gel using dedicated software usually provided with the camera. First, the position of the cursor relative to the DNA band is adjusted to achieve optimum size matching with the reference strain used as internal control, and then the cursor is similarly positioned for all the other strains run in the same gel. Size assign-

ment is confirmed by visual inspection of gels and comparative analysis of strains for each marker. For repeats of 12 bp and more, visual estimation of the band size is possible with the help of a chart in which all the known alleles are indicated (*see* **Note 9**).

3.1.5. Capillary Gel Electrophoresis

Apart from the regular agarose gel, the equipment most often used is capillary electrophoresis apparatus, in particular equipment initially developed for DNA sequencing purposes. The precision and reproducibility achieved by such machines (routinely ±0.5 bp in a 80- to 650-bp range) enables the typing of very short repeat units, although the typing of long mononucleotides or even dinucleotide repeat tracts can be technically demanding *(6)*. Also the size estimates provided, deduced by comparison with a size standard, can be wrong by a few basepairs in a very reproducible way. This has been illustrated by different studies *(7,8)*. The discrepancy for a given machine and for each locus and allele must be measured experimentally. These machines require the use of fluorescent primers. Because the underlying technology was developed for DNA sequencing purposes, at least four different colors are available, one of which must be used for the DNA size standard. This availability of different colors makes it possible to pool different PCR products to analyze multiple loci in the same run (*see* **Note10**) and consequently reduce the costs.

Other capillary electrophoresis machines specifically developed for measuring the length of DNA fragments (rather than sequencing) and that do not require the use of fluorescent primers are also very promising (including Agilent 2100, Caliper Labchip90, Qiagen QIAxcel), in spite of a slightly lower precision *(7,8)*.

3.1.6. Nomenclature and Description of MLVA Profiles

The repeat length and number of repetitions are conveniently determined in sequenced genomes using the Microbial Tandem Repeats Database (http://minisatellites.u-psud.fr) *(9,10)*.

1. Check that amplification of DNA from the strain from which the genome has been sequenced produces amplicons of the expected size.
2. Estimate the number of repeats in new alleles by subtracting the invariable flanking region from the amplicon size, then dividing by the repeat unit length as determined for the reference strain. For example, if the size of the PCR amplification product for a 45-bp VNTR is 205 bp in the sequenced strain for two repeat units, it implies that the number of repeats found in amplicons of size x produced with the same primers is $((x - 205)/45 + 2)$ repeat units.
3. The null designation is given when no amplification is repeatedly observed at a given locus (*see* **Note 11**).

3.1.7. Verification of Unexpected Allele Size

Intermediate-size alleles may result from intermediate-size repeat units or from small deletions in the flanking sequence. Sequencing

of any such allele is necessary to analyze the origin of the unexpected size.

1. The full-length sequences of selected PCR products are determined on both strands following DNA purification by the QIAquick PCR purification kit (Qiagen) or by polyethylene glycol (PEG) precipitation as described in **ref.** *11*. Data obtained with forward- and reverse-sequencing primers are combined, and sequences are manually aligned.
2. The alleles are reported as half size to indicate the existence of abnormal alleles.

3.1.8. Data Storage and Analysis

Ideally, data should be stored in databases together with all known information on the strain (i.e., phenotypic and biochemical characteristics, origin, clinical or environmental information, etc.).

1. For each isolate, enter the VNTR size in the form of a digit corresponding to the number of repeats into an Excel file.
2. Import the data matrix into data-mining tools or into more conventional biology-oriented clustering methods. The currently preferred method to measure the distance between two strains is simple counting of the number of markers at which the two strains differ divided by the total number of markers and expressed as a percentage (*see* **Note 12**).
3. When the amount of data is sufficient, it becomes possible to precisely estimate the mutation rate at each VNTR and the relative frequency of single and multiple repeat unit gains and losses at each locus *(12–14)*. Using this knowledge, it is possible to define individual distance coefficients (*see* **Note 13**).
4. Once a good-quality data set has been produced, it is desirable to make it accessible via the Internet (*see* **Note 14**). At least three research groups (*see* **Note 15**) have developed sites that make possible the hosting of databases produced by other groups. For instance, using the MLVAbank at http://mlva.u-psud.fr, MLVA typers can create their own account to manage personal databases without control from the hosting institution. Databases can be made public, or be shared within a community, with different rights. Different panels of markers can be selected by users to take into account local usages. This hosting facility may be seen as a repository for MLVA data. It is not curated by the hosting institution in an approach that is reminiscent of DNA sequence repositories such as Genbank and European Molecular Biology Laboratory (EMBL).

3.2. Developing a New MLVA Assay

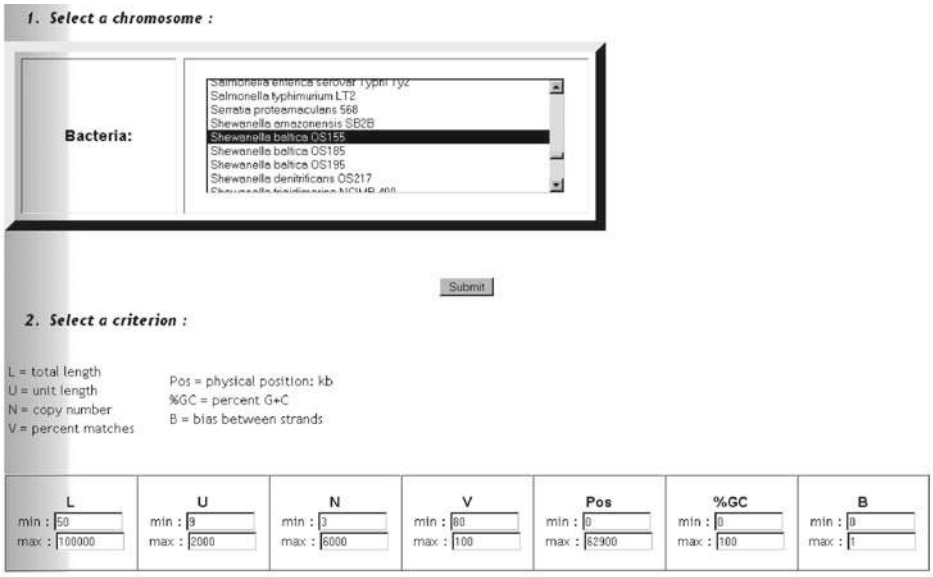
Once a bacterial genome has been sequenced (fully or partially), it is possible to search for tandem repeats to test the potential use of MLVA for genotyping of the species. The availability of two or

more strains of the species highly facilitates the search for polymorphic markers. Web-based tools have been developed to facilitate the first steps in setting a new MLVA scheme (*see* **Note 16**).

3.2.1. Identification of Tandem repeats from a Single Genome

1. Go to the Microorganisms Tandem Repeat Database (http://minisatellites.u-psud.fr/) developed by Denoeud and Vergnaud ***(10)*** and follow the link "The Microorganisms Tandem Repeats Database" and then "bacteria". Select the strain to be searched and choose the parameters that will define the tandem repeats of interest. For a first assay, it is interesting to select repeats of 9 bp and longer, repeated at least three times and showing an 80% internal conservation. When repeats fulfilling these criteria exist, they are displayed in a table as shown on **Fig. 2**.
2. Click on the alignment link to see the repeat sequence and a 500-bp sequence flanking it on both sides.

a



b

Physical position (kb)	Contig name	Unit length	Copy number	Total length	Percent matches	%C	%G	%T	%A	GC bias	Sequence	TR of the family
4	Shewanella_baltica_OS155	9	13.8	123	80%	6%	38%	11%	43%	0.72	alignment / linear seq	Group
12	Shewanella_baltica_OS155	261	12.1	3161	88%	30%	24%	23%	21%	0.11	alignment / linear seq	Group
29	Shewanella_baltica_OS155	24	3.0	73	100%	8%	24%	42%	24%	0.49	alignment / linear seq	Group
31	Shewanella_baltica_OS155	9	5.9	53	93%	22%	9%	24%	44%	0.41	alignment / linear seq	Group
102	Shewanella_baltica_OS155	18	7.8	140	91%	17%	44%	12%	25%	0.44	alignment / linear seq	Group
623	Shewanella_baltica_OS155	118	3.7	436	98%	20%	20%	28%	29%	0	alignment / linear seq	Group

Fig. 2. The tandem repeats database. Snapshots of the Web-based database query and of the output. (**a**) The strain of interest is selected as well as the characteristics of the VNTRs to be searched. (**b**) The repeats are displayed in a table showing the position on the genome, the repeat size, and copy number as well as other information on its composition. The sequence can be retrieved using the link in the rightmost column.

3. Select primers in sequences flanking the VNTR, at least 40 bp away from the first and last repeats (*see* **Note 17**), in such a way that the same annealing temperature can be used for all the PCRs. This is particularly important if multiplexing is to be used.

3.2.2. Identification of VNTR by Comparison of Two or More Genomes

When the genomes of two strains of a given species have been sequenced, it is possible to compare the size of repeated sequences to select those possessing a different number of repeats and therefore representing good VNTR candidates.

1. Access the strain comparison page (from http://minisatellites.u-psud.fr/, follow the link "The Microorganisms Tandem Repeats Database" and then "strain comparison page") and select the bacteria to be searched.
2. Choose the tandem repeats characteristics.
3. A table will show repeats that present polymorphism between the two strains.

3.2.3. The Choice of a Panel of Markers

The excessive use of microsatellites (2- to 8-bp repeat units), some of which tend to be unstable, as reported in several studies, may cause difficulties because of their especially high homoplasy levels ***(3)***. In addition, they necessitate the use of sequencing gels or methods with equivalent resolution, which are usually not routinely used in the bacteriology laboratory.

3.2.4. Criteria for Evaluation of MLVA

Standard efficacy criteria of a new MLVA scheme, including typeability (*T*), reproducibility (*R*), stability (*S*), epidemiological concordance (*E*), and discriminatory power (expressed as the Hunter-Gaston diversity index [*HGDI*]), are determined as reported elsewhere ***(15,16)***.

The polymorphism index of individual or combined VNTR loci can be calculated using a selected panel of strains and *HGDI* ***(17)***, an application of the Simpson' s index of diversity ***(18)***. Although very useful to compare the discriminatory power of assays, it does not measure the relevance of the discrimination that is achieved by a given marker or combination of markers (*see* **Note 18**).

3.3. Reviewing MLVA Reports

MLVA typing is still an emerging domain, and the quality of MLVA reports is unequal. In particular, some reports neglect basic rules, some of which are specific for MLVA. To accelerate the development of MLVA, we propose here a checklist for reviewers of MLVA articles:

Check that the proposed markers are indeed new markers and have not been previously described under other names.

Check that the MLVA typing data are made accessible and that the allele-calling convention is clearly described by giving the repeat unit size and repeat copy number in the first genome

sequenced for the corresponding species. Encourage the deposition of data into one of the existing MLVA data repositories.

Check that the assay is an MLVA assay and not a pattern-based assay in which multiple loci are simultaneously revealed, but not analyzed to the point of deducing the repeat copy number at each locus individually.

3.4. Future Prospects

In spite of very promising progress and significant advantages, at least in theory, MLVA is not replacing existing technologies, such as pulsed-field gel electrophoresis, as fast as could have been expected. There are several reasons for this. One is the lack of standardization and reference databases. In this area, funding bodies have a major role to play by stimulating and supporting the actors involved to agree on international standards. Another reason is the lack of demand for large-scale molecular epidemiological tools. The United States are one exception illustrating this fact: a large unified market has led to requirements for genotyping systems covering the whole country. No such requirements exist in Europe, where national-scale approaches are still the rule in practice. Usually, a single national reference laboratory will organize molecular typing, and in this context, pattern-based approaches can be used in spite of their limitations (limited possibilities of interlaboratory exchanges). Still another reason for the slow emergence of MLVA is its relative cost as it requires multiple PCR amplifications. Once reference typing panels have been defined, it is hoped that multiplexing technologies will be developed to permit genotyping of a strain using (ideally) a single PCR amplification.

4. Notes

1. From a technical point of view, it is convenient to classify tandem repeats among three classes: the "minisatellites" with repeat units longer than 9 bp, the microsatellites with repeat units in the 2- to 8-bp range, and the homopolymeric repeats (often called Single Nucleotide Repeats or SNRs). The larger repeat unit loci can usually be typed on a wide range of DNA fragment-sizing equipment. The microsatellites will usually require more sophisticated equipment, and the mononucleotide repeats with 1-bp variations require specific protocols, including the use of different polymerases, such as *Pfu* ***(6)***.
2. The simple addition of betain is very effective in terms of priming specificity. It may help resolve multiple-band amplifications. To have its full effect, the use of a special PCR buffer is recommended ***(19)***.

3. Size marker: Select size markers containing similar amounts of DNA in each band so that the fluorescence intensity of each band is similar.
4. A voltage of 8–10 V/cm is usually applied (i.e., if the distance between the electrodes is 30 cm, voltage is 250–300 V). If a cover is necessary, it may be important to check for buffer temperature and avoid overheating by recirculating and cooling the buffer. If the two electrodes are identical (i.e., platinum electrodes), it is recommended to revert the migration polarity every five runs to avoid running distortions.
5. VNTR panel consensus: Special effort should be put to reach a consensus on which panel of VNTR to use for a given MLVA assay. This is being achieved for *Brucella*, for which a set of different panels with increasing discriminatory power has been defined.
6. Shorter repeat units or larger size ranges can be resolved using higher resolution, usually with precast gels, such as Biorex.
7. Gel quality: Special attention must be given to the quality of the gel as this will allow precise size assignment. It is recommended to pour the gel at a temperature between 60 and 65°C.
8. Ethidium bromide staining: It is mandatory to stain the gel after migration, especially when looking at small-size PCR product as the ethidium bromide in the gel will migrate backward.
9. See, for instance, http://mlva.u-psud.fr/BRUCELLA/spip.php?article93.
10. The development of this equipment was driven by sequencing, and DNA fragment size measurement is only a secondary application. With the advent of new sequencing technologies, it remains to be seen whether this "DNA fragment-sizing" market will be sufficient to maintain these machines or if other approaches will be needed. A number of alternative equipment, such as the ones developed or distributed for instance by QIAGEN, Caliper, and Agilent, might then replace capillary apparatus aimed at DNA sequencing.
11. In some instances, an inappropriate initial numbering may result in the calling of "zero-repeat" alleles as more strains are investigated ***(20)***. This is because very often a tandem repeat locus does not contain an integer number of repeats but rather contains a partial repeat at one end. If alleles at such a locus with, for instance, 1.5, 2.5, 3.5 repeat units are called 1, 2, 3 rather than 2, 3, 4, an allele containing 0.5 repeats will be coded 0. Eventually, normalizing bodies will be needed to avoid such ambiguities.
12. This is a very crude similarity measure that gives the same weight to all markers. It also considers that alleles that differ by one repeat unit are not evolutionarily closer than alleles that differ by many repeat units. The two assumptions are often wrong,

but in spite of this, the resulting clustering analyses make sense. This is because the use of multiple markers compensates for variable homoplasy levels at individual markers.

13. *Brucella* MLVA data, for instance, are already analyzed by dividing the markers into three different sets, or panels, and giving a different weight to each panel.
14. An MLVA data set was made accessible in an interactive way for the first time in 2002 ***(21)***. Since then, a few other databases dedicated to one or a few pathogens have been put on line.
15. See http://mlva.u-psud.fr; http://www.mlva.eu; http://www.pasteur.fr/mlva.
16. The development of MLVA usually comprises three phases. In the first phase, polymorphic markers are identified. Usually, a few publications cover this first step. In the second phase, more typing data are produced, and the characteristics of individual markers are refined. Consensus marker panels progressively emerge. In the third phase, typing databases are produced, and consensus panels are agreed on.
17. In this way, representative alleles can more easily be sequenced using the same primers. Otherwise, if the primers are located too close from the tandem repeat start, sequencing data will often miss the first basepairs of the tandem repeat.
18. Eventually, it should make sense to consider that two strains that differ at one highly variable marker are more similar than two strains that differ at a moderately variable marker. More sophisticated distance coefficients can be developed once many strains have been typed.

Acknowledgment

We thank the CNRS and Université Paris Sud 11.

References

1. Vergnaud, G., and Denoeud, F. (2000). Minisatellites: mutability and genome architecture. *Genome Res.* **10**, 899–907.
2. Vergnaud, G., and Pourcel, C. (2006). Multiple locus VNTR (variable number of tandem repeat) analysis, in *Molecular Identification, Systematics, and Population Structure of Prokaryotes* (Stackebrandt, E., ed.), Springer-Verlag, Berlin, pp. 83–104.
3. Bricker, B. J., Ewalt, D. R., and Halling, S. M. (2003). *Brucella* "HOOF-Prints": strain typing by multi-locus analysis of variable number tandem repeats (VNTRs). *BMC Microbiol.* **3**, 15.
4. Laroucau, K., Thierry, S., Vorimore, F., Blanco, K., Kaleta, E., Hoop, R., et al. (2008). High resolution typing of *Chlamydophila psittaci* by multilocus VNTR analysis (MLVA). *Infect. Genet. Evol.* **8**, 171–181.
5. Vu-Thien, H., Corbineau, G., Hormigos, K., Fauroux, B., Corvol, H., Clement, A., et al. (2007). Multiple-locus variable-number

tandem-repeat analysis for longitudinal survey of sources of *Pseudomonas aeruginosa* infection in cystic fibrosis patients. *J. Clin. Microbiol.* **45**, 3175–3183.

6. Kenefic, L. J., Beaudry, J., Trim, C., Huynh, L., Zanecki, S., Matthews, M., et al. (2008). A high resolution four-locus multiplex single nucleotide repeat (SNR) genotyping system in *Bacillus anthracis*. *J. Microbiol. Methods* **73**, 269–272.
7. Lista, F., Faggioni, G., Valjevac, S., Ciammaruconi, A., Vaissaire, J., le Doujet, C., et al. (2006). Genotyping of *Bacillus anthracis* strains based on automated capillary 25-loci multiple locus variable-number tandem repeats analysis. *BMC Microbiol.* **6**, 33.
8. Ciammaruconi, A., Grassi, S., De Santis, R., Faggioni, G., Pittiglio, V., D'Amelio, R., et al. (2008). Fieldable genotyping of *Bacillus anthracis* and *Yersinia pestis* based on 25-loci multi locus VNTR analysis. *BMC Microbiol.* **8**, 21.
9. Le Flèche, P., Hauck, Y., Onteniente, L., Prieur, A., Denoeud, F., Ramisse, V., et al. (2001). A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*. *BMC Microbiol.* **1**, 2.
10. Denoeud, F., and Vergnaud, G. (2004). Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains: a Web-based resource. *BMC Bioinformatics* **5**, 4.
11. Embley, T. M. (1991). The linear PCR reaction: a simple and robust method for sequencing amplified rRNA genes. *Lett. Appl. Microbiol.* **13**, 171–174.
12. U'Ren, J. M., Schupp, J. M., Pearson, T., Hornstra, H., Friedman, C. L., Smith, K. L., et al. (2007). Tandem repeat regions within the *Burkholderia pseudomallei* genome and their application for high resolution genotyping. *BMC Microbiol.* **7**, 23.
13. Whatmore, A. M., Shankster, S. J., Perrett, L. L., Murphy, T. J., Brew, S. D., Thirlwall, R. E., et al. (2006). Identification and characterization of variable-number tandem-repeat markers for typing of *Brucella* spp. *J. Clin. Microbiol.* **44**, 1982–1993.
14. Garcia-Yoldi, D., Le Fleche, P., De Miguel, M. J., Munoz, P. M., Blasco, J. M., Cvetnic, Z., et al. (2007). Comparison of multiple-locus variable-number tandem-repeat analysis with other PCR-based methods for typing *Brucella suis* isolates. *J. Clin. Microbiol.* **45**, 4070–4072.
15. Struelens, M. J. (1996). Consensus guidelines for appropriate use and evaluation of microbial epidemiologic typing systems. *Clin. Microbiol. Infect.* **2**, 2–11.
16. Pourcel, C., Visca, P., Afshar, B., D'Arezzo, S., Vergnaud, G., and Fry, N. K. (2007). Identification of variable-number tandem-repeat (VNTR) sequences in *Legionella pneumophila* and development of an optimized multiple-locus VNTR analysis typing scheme. *J. Clin. Microbiol.* **45**, 1190–1199.
17. Hunter, P. R., and Gaston, M. A. (1988). Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. *J. Clin. Microbiol.* **26**, 2465–2466.
18. Simpson, E. H. (1949). Measurement of diversity. *Nature*, **163**, 688.
19. Henke, W., Herdel, K., Jung, K., Schnorr, D., and Loening, S. A. (1997). Betaine improves the PCR amplification of GC-rich DNA sequences. *Nucleic Acids Res.* **25**, 3957–3958.
20. Fabre, M., Koeck, J. L., Le Fleche, P., Simon, F., Herve, V., Vergnaud, G., et al. (2004). High genetic diversity revealed by variable-number tandem repeat genotyping and analysis of hsp65 gene polymorphism in a large collection of "*Mycobacterium canettii*" strains indicates that the *M. tuberculosis* complex is a recently emerged clone of "*M. canettii*." *J. Clin. Microbiol.* **42**, 3248–3255.
21. Le Flèche, P., Fabre, M., Denoeud, F., Koeck, J. L., and Vergnaud, G. (2002). High resolution, on-line identification of strains from the *Mycobacterium tuberculosis* complex based on tandem repeat typing. *BMC Microbiol.* **2**, 37.
22. Keim, P., Price, L. B., Klevytska, A. M., Smith, K. L., Schupp, J. M., Okinaka, R., et al. (2000). Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis*. *J. Bacteriol.* **182**, 2928–2936.
23. Monteil, M., Durand, B., Bouchouicha, R., Petit, E., Chomel, B., Arvand, M., et al. (2007). Development of discriminatory multiple-locus variable number tandem repeat analysis for *Bartonella henselae*. *Microbiology* **153**, 1141–1148.
24. Schouls, L. M., van der Heide, H. G., Vauterin, L., Vauterin, P., and Mooi, F. R. (2004). Multiple-locus variable-number tandem repeat analysis of Dutch *Bordetella pertussis* strains reveals rapid genetic changes with clonal expansion during the late 1990s. *J. Bacteriol.* **186**, 5496–5505.
25. Farlow, J., Postic, D., Smith, K. L., Jay, Z., Baranton, G., and Keim, P. (2002). Strain typing of *Borrelia burgdorferi*, *Borrelia afzelii*, and *Borrelia garinii* by using multiple-locus

variable-number tandem repeat analysis. *J. Clin. Microbiol.* **40**, 4612–4618.

26. Le Fleche, P., Jacques, I., Grayon, M., Al Dahouk, S., Bouchon, P., Denoeud, F., et al. (2006). Evaluation and selection of tandem repeat loci for a *Brucella* MLVA typing assay. *BMC Microbiol.* **6**, 9.
27. Botterel, F., Desterke, C., Costa, C., and Bretagne, S. (2001). Analysis of microsatellite markers of *Candida albicans* used for rapid typing. *J. Clin. Microbiol.* **39**, 4076–4081.
28. Grenouillet, F., Millon, L., Bart, J. M., Roussel, S., Biot, I., Didier, E., et al. (2007). Multiple-locus variable-number tandem-repeat analysis for rapid typing of *Candida glabrata*. *J. Clin. Microbiol.* **45**, 3781–3784.
29. van den Berg, R. J., Schaap, I., Templeton, K. E., Klaassen, C. H., and Kuijper, E. J. (2007). Typing and subtyping of *Clostridium difficile* isolates by using multiple-locus variable-number tandem-repeat analysis. *J. Clin. Microbiol.* **45**, 1024–1028.
30. Marsh, J. W., O'Leary, M. M., Shutt, K. A., Pasculle, A. W., Johnson, S., Gerding, D. N., et al. (2006). Multilocus variable-number tandem-repeat analysis for investigation of *Clostridium difficile* transmission in hospitals. *J. Clin. Microbiol.* **44**, 2558–2566.
31. Sawires, Y. S., and Songer, J. G. (2005). Multiple-locus variable-number tandem repeat analysis for strain typing of *Clostridium perfringens*. *Anaerobe* **11**, 262–272.
32. Arricau-Bouvery, N., Hauck, Y., Bejaoui, A., Frangoulidis, D., Bodier, C. C., Souriau, A., et al. (2006). Molecular characterization of *Coxiella burnetii* isolates by infrequent restriction site-PCR and MLVA typing. *BMC Microbiol.* **6**, 38.
33. Svraka, S., Toman, R., Skultety, L., Slaba, K., and Homan, W. L. (2006). Establishment of a genotyping scheme for *Coxiella burnetii*. *FEMS Microbiol. Lett.* **254**, 268–274.
34. Titze-de-Almeida, R., Willems, R.J., Top, J., Rodrigues, I. P., Ferreira, R. F.2nd, Boelens, H., et al. (2004). Multilocus variable-number tandem-repeat polymorphism among Brazilian *Enterococcus faecalis* strains. *J. Clin. Microbiol.* **42**, 4879–4881.
35. Top, J., Schouls, L. M., Bonten, M. J., and Willems, R. J. (2004). Multiple-locus variable-number tandem repeat analysis, a novel typing scheme to study the genetic relatedness and epidemiology of *Enterococcus faecium* isolates. *J. Clin. Microbiol.* **42**, 4503–4511.
36. Noller, A. C., McEllistrem, M. C., Pacheco, A. G., Boxrud, D. J., and Harrison, L. H. (2003). Multilocus variable-number tandem repeat analysis distinguishes outbreak and sporadic *Escherichia coli* O157:H7 isolates. *J. Clin. Microbiol.* **41**, 5389–5397.
37. Lindstedt, B. A., Heir, E., Gjernes, E., Vardund, T., and Kapperud, G. (2003). DNA fingerprinting of Shiga-toxin producing *Escherichia coli* O157 based on multiple-locus variable-number tandem-repeats analysis (MLVA). *Ann. Clin. Microbiol. Antimicrob.* **2**, 12.
38. Lindstedt, B. A., Brandal, L. T., Aas, L., Vardund, T., and Kapperud, G. (2007). Study of polymorphic variable-number of tandem repeats loci in the ECOR collection and in a set of pathogenic *Escherichia coli* and *Shigella* isolates for use in a genotyping assay. *J. Microbiol. Methods* **69**, 197–205.
39. Farlow, J., Smith, K. L., Wong, J., Abrams, M., Lytle, M., and Keim, P. (2001). *Francisella tularensis* strain typing using multiple-locus, variable-number tandem repeat analysis. *J. Clin. Microbiol.* **39**, 3186–3192.
40. Johansson, A., Forsman, M., and Sjostedt, A. (2004). The development of tools for diagnosis of tularemia and typing of *Francisella tularensis*. *APMIS* **112**, 898–907.
41. van Belkum, A., Scherer, S., van Leeuwen, W., Willemse, D., van Alphen, L., and Verbrugh, H. (1997). Variable number of tandem repeats in clinical strains of *Haemophilus influenzae*. *Infect. Immun.* **65**, 5017–5027.
42. Diancourt, L., Passet, V., Chervaux, C., Garault, P., Smokvina, T., and Brisse, S. (2007). Multilocus sequence typing of *Lactobacillus casei* reveals a clonal population structure with low levels of homologous recombination. *Appl. Environ. Microbiol.* **73**, 6601–6611.
43. Pourcel, C., Vidgop, Y., Ramisse, F., Vergnaud, G., and Tram, C. (2003). Characterization of a tandem repeat polymorphism in *Legionella pneumophila* and its use for genotyping. *J. Clin. Microbiol.* **41**, 1819–1826.
44. Majed, Z., Bellenger, E., Postic, D., Pourcel, C., Baranton, G., and Picardeau, M. (2005). Identification of variable-number tandem-repeat loci in *Leptospira interrogans* sensu stricto. *J. Clin. Microbiol.* **43**, 539–545.
45. Slack, A. T., Dohnt, M. F., Symonds, M. L., and Smythe, L. D. (2005). Development of a multiple-locus variable number of tandem repeat analysis (MLVA) for *Leptospira interrogans* and its application to *Leptospira interrogans* serovar Australis isolates from Far North Queensland, Australia. *Ann. Clin. Microbiol. Antimicrob.* **4**, 10.
46. Salaun, L., Merien, F., Gurianova, S., Baranton, G., and Picardeau, M. (2006). Application of multilocus variable-number tandem-repeat analysis for molecular typing of the agent of leptospirosis. *J. Clin. Microbiol.* **44**, 3954–3962.

47. Murphy, M., Corcoran, D., Buckley, J. F., O'Mahony, M., Whyte, P., and Fanning, S. (2007). Development and application of multiple-locus variable number of tandem repeat analysis (MLVA) to subtype a collection of *Listeria monocytogenes*. *Int. J. Food Microbiol.* **115**, 187–194.
48. Bull, T. J., Sidi-Boumedine, K., McMinn, E. J., Stevenson, K., Pickup, R., and Hermon-Taylor, J. (2003). Mycobacterial interspersed repetitive units (MIRU) differentiate *Mycobacterium avium* subspecies paratuberculosis from other species of the *Mycobacterium avium* complex. *Mol. Cell Probes* **17**, 157–164.
49. Overduin, P., Schouls, L., Roholl, P., van der Zanden, A., Mahmmod, N., Herrewegh, A., et al. (2004). Use of multilocus variable-number tandem-repeat analysis for typing *Mycobacterium avium* subsp. paratuberculosis. *J. Clin. Microbiol.* **42**, 5022–5028.
50. Truman, R., Fontes, A. B., De Miranda, A. B., Suffys, P., and Gillis, T. (2004). Genotypic variation and stability of four variable-number tandem repeats and their suitability for discriminating strains of *Mycobacterium leprae*. *J. Clin. Microbiol.* **42**, 2558–2565.
51. Groathouse, N. A., Rivoire, B., Kim, H., Lee, H., Cho, S. N., Brennan, P. J., et al. (2004). Multiple polymorphic loci for molecular typing of strains of *Mycobacterium leprae*. *J. Clin. Microbiol.* **42**, 1666–1672.
52. Frothingham, R., and Meeker-O'Connell, W.A. (1998). Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats. *Microbiology* **144**, 1189–1196.
53. Supply, P., Mazars, E., Lesjean, S., Vincent, V., Gicquel, B., and Locht, C. (2000). Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. *Mol. Microbiol.* **36**, 762–771.
54. Skuce, R. A., McCorry, T. P., McCarroll, J. F., Roring, S. M., Scott, A. N., Brittain, D., et al. (2002). Discrimination of *Mycobacterium tuberculosis* complex bacteria using novel VNTR-PCR targets. *Microbiology* **148**, 519–528.
55. Ablordey, A., Swings, J., Hubans, C., Chemlal, K., Locht, C., Portaels, F., et al. (2005). Multilocus variable-number tandem repeat typing of *Mycobacterium ulcerans*. *J. Clin. Microbiol.* **43**, 1546–1551.
56. Stragier, P., Ablordey, A., Meyers, W. M., and Portaels, F. (2005). Genotyping *Mycobacterium ulcerans* and *Mycobacterium marinum* by using mycobacterial interspersed repetitive units. *J. Bacteriol.* **187**, 1639–1647.
57. McAuliffe, L., Ayling, R. D., and Nicholas, R. A. (2007). Identification and characterization of variable-number tandem-repeat markers for the molecular epidemiological analysis of *Mycoplasma mycoides* subspecies *mycoides* SC. *FEMS Microbiol. Lett.* **276**, 181–188.
58. Liao, J. C., Li, C. C., and Chiou, C. S. (2006). Use of a multilocus variable-number tandem repeat analysis method for molecular subtyping and phylogenetic analysis of *Neisseria meningitidis* isolates. *BMC Microbiol.* **6**, 44.
59. Schouls, L. M., van der Ende, A., Damen, M., and van de Pol, I. (2006). Multiple-locus variable-number tandem repeat analysis of *Neisseria meningitidis* yields groupings similar to those obtained by multilocus sequence typing. *J. Clin. Microbiol.* **44**, 1509–1518.
60. Onteniente, L., Brisse, S., Tassios, P. T., and Vergnaud, G. (2003). Evaluation of the polymorphisms associated with tandem repeats for *Pseudomonas aeruginosa* strain typing. *J. Clin. Microbiol.* **41**, 4991–4997.
61. Lindstedt, B. A., Heir, E., Gjernes, E., and Kapperud, G. (2003). DNA fingerprinting of *Salmonella enterica* subsp. *enterica* serovar typhimurium with emphasis on phage type DT104 based on variable number of tandem repeat loci. *J. Clin. Microbiol.* **41**, 1469–1479.
62. Ramisse, V., Houssu, P., Hernandez, E., Denoeud, F., Hilaire, V., Lisanti, O., et al. (2004). Variable number of tandem repeats in *Salmonella enterica* subsp. *enterica for typing purposes*. *J. Clin. Microbiol.* **42**, 5722–5730.
63. Hardy, K. J., Ussery, D. W., Oppenheim, B. A., and Hawkey, P. M. (2004). Distribution and characterization of staphylococcal interspersed repeat units (SIRUs) and potential use for strain differentiation. *Microbiology* **150**, 4045–4052.
64. Sabat, A., Krzyszton-Russjan, J., Strzalka, W., Filipek, R., Kosowska, K., Hryniewicz, W., et al. (2003). New method for typing *Staphylococcus aureus* strains: multiple-locus variable-number tandem repeat analysis of polymorphism and genetic relationships of clinical isolates. *J. Clin. Microbiol.* **41**, 1801–1804.
65. Francois, P., Huyghe, A., Charbonnier, Y., Bento, M., Herzig, S., Topolski, I., et al. (2005). Use of an automated multiple-locus, variable-number tandem repeat-based method for rapid and high-throughput genotyping of *Staphylococcus aureus* isolates. *J. Clin. Microbiol.* **43**, 3346–3355.
66. Koeck, J. L., Njanpop-Lafourcade, B. M., Cade, S., Varon, E., Sangare, L., Valjevac, S., et al. (2005). Evaluation and selection of tandem repeat loci for *Streptococcus pneumoniae* MLVA strain typing. *BMC Microbiol.* **5**, 66.
67. Gilbert, F. B., Fromageau, A., Lamoureux, J., and Poutrel, B. (2006). Evaluation of tandem

repeats for MLVA typing of *Streptococcus uberis* isolated from bovine mastitis. *BMC Vet. Res.* **2**, 33.

68. Witonski, D., Stefanova, R., Ranganathan, A., Schutze, G.E., Eisenach, K. D., and Cave, M. D. (2006). Variable-number tandem repeats that are useful in genotyping isolates of *Salmonella enterica* subsp. *enterica* serovars Typhimurium and Newport. *J. Clin. Microbiol.* **44**, 3849–3854.
69. Liang, S. Y., Watanabe, H., Terajima, J., Li, C. C., Liao, J. C., Tung, S. K., et al. (2007). Multilocus variable-number tandem-repeat analysis for molecular typing of *Shigella sonnei*. *J. Clin. Microbiol.* **45**, 3574–3580.
70. Oura, C. A., Odongo, D. O., Lubega, G. W., Spooner, P. R., Tait, A., and Bishop, R. P. (2003). A panel of microsatellite and minisatellite markers for the characterisation of field isolates of *Theileria parva*. *Int. J. Parasitol.* **33**, 1641–1653.
71. Coletta-Filho, H. D., Takita, M. A., de Souza, A. A., Aguilar-Vildoso, C. I., and Machado, M. A. (2001). Differentiation of strains of *Xylella fastidiosa* by a variable number of tandem repeat analysis. *Appl. Environ. Microbiol.* **67**, 4091–4095.
72. Pourcel, C., Andre-Mazeaud, F., Neubauer, H., Ramisse, F., and Vergnaud, G. (2004). Tandem repeats analysis for the high resolution phylogenetic analysis of *Yersinia pestis*. *BMC Microbiol.* **4**, 22.
73. Klevytska, A. M., Price, L. B., Schupp, J. M., Worsham, P. L., Wong, J., and Keim, P. (2001). Identification and characterization of variable-number tandem repeats in the *Yersinia pestis* genome. *J. Clin. Microbiol.* **39**, 3179–3185.
74. Achtman, M., Morelli, G., Zhu, P., Wirth, T., Diehl, I., Kusecek, B., et al. (2004). Microevolution and history of the plague bacillus, *Yersinia pestis*. *Proc. Natl. Acad. Sci. USA* **101**, 17837–17842.

Chapter 13

Comparison of Molecular Typing Methods Applied to *Clostridium difficile*

Ed J. Kuijper, Renate J. van den Berg, and Jon S. Brazier

Abstract

Since the 1980s the epidemiology of *Clostridium difficile* infection (CDI) has been investigated by the application of many different typing or fingerprinting methods. To study the epidemiology of CDI, a typing method with a high discriminatory power, typeability, and reproducibility is required. Molecular typing methods are generally regarded as having advantages over phenotypic methods in terms of the stability of genomic markers and providing greater levels of typeability. A growing number of molecular methods have been applied to *C. difficile*. For the early and rapid detection of outbreak situations, methods such as restriction enzyme analysis, arbitrary primed polymerase chain reaction (PCR), and PCR ribotyping are commonly used. For long-term epidemiology, multilocus sequence typing, multilocus variable number of tandem repeats analysis, and amplified fragment length polymorphism are of interest. Currently, the PCR-ribotyping method and the library of PCR ribotypes in Cardiff are the benchmarks to which most typing studies around the world are compared. Multilocus variable number of tandem repeats analysis is the most discriminative typing method and will contribute significantly to our understanding of the epidemiology of this important nosocomial pathogen.

Key words: *Clostridium difficile*, MLVA, PCR ribotyping, REA, subtyping.

1. Introduction

Since the recognition of *Clostridium difficile* as the causative agent of pseudomembranous colitis in 1978, this anaerobic spore-forming bacterium has emerged as an important enteropathogen. Pathogenic *C. difficile* organisms release toxins that ultimately mediate diarrhea and colitis. Colonic injury and inflammation result from the production of two protein toxins: enterotoxin A (TcdA; 308,000 M_r) and cytotoxin B (TcdB; 270,000 M_r). Genes for the

D.A. Caugant (ed.), *Molecular Epidemiology of Microorganisms*, Methods in Molecular Biology, Vol. 551
DOI: 10.1007/978-1-60327-999-4_13,

Table 1
Characteristics of Genotyping Methods

Method	Target	Discriminatory power	Typeability	Reproducibility	Performance	Interpretation	Costs	Interlaboratory exchange
Plasmid profiling	Extrachromosomal plasmid	–	–	+	±	±	+	+
REA	Whole genome, restriction	+	+	±	±	–	+	–
RFLP	Whole genome, restriction	–	+	±	±	±	+	–
AP-PCR/RAPD	Whole genome, random PCR primers	++	+	±	+	+	+	–
PCR ribotyping	16S–23S intergenic spacer region	++	+	+++	+++	++	+	+++
PFGE	Whole genome, restriction	+++	±	+++	±	±	+++	±
Toxinotyping	Toxin A and B genes	+	+	+++	++	+	+	++
fliC PCR-RFLP	Flagellin gene	±	+	++	++	++	+	++
slpA PCR-RFLP	S-layer precursor gene	±	+	++	++	++	+	++
AFLP	Whole genome, restriction	++	+	++	±	+	++	±
MLST	Seven housekeeping and ten virulence-associated genes	++	+	+++	++	+++	++	+++
MLVA	Whole genome, tandem repeats	+++	+	+++	++	+++	++	+++

REA; restriction enzyme analysis, RFLP; restriction fragment length polymorphism, AP-PCR; arbitrary primed PCR, RAPD; random amplified polymorphic DNA, PFGE; pulsed-field gel electrophoresis, AFLP; amplified fragment length polymorphism, MLST; multilocus sequence typing, MLVA; multilocus VNTR analysis

binary toxin are located outside the pathogenicity locus (PaLoc), but the role of this toxin is unclear *(1)*. The illness associated with *C. difficile* (*C. difficile* infection, CDI) ranges from mild diarrhea to life-threatening colitis.

To study the epidemiology of CDI, a typing method with a high discriminatory power, typeability, and reproducibility is required. Typing methods are also used to determine the role of the environment and patient-to-patient transmission in the cause of infection and for the investigation of outbreaks. The recurrence rate of *C. difficile*-associated disease is around 15–20%, and typing methods can be applied to distinguish recurrences in relapse, due to the same strain, or reinfection, due to a new strain *(1)*.

Typing methods can be classified in two large categories, consisting of phenotypic and genotypic methods. Phenotypic methods differentiate on the basis of products of gene expression, whereas genotyping methods analyze the genetic profile of the strains. Molecular typing methods are generally regarded as having advantages over phenotypic methods in terms of the stability of genomic markers and providing greater levels of typeability. A growing number of molecular methods have been applied to *C. difficile*, and these are described here (*see* **Table 1** for an overview).

2. Traditional Molecular Typing Methods for *C. difficile*

2.1. Plasmid profiling

Plasmid profiling was the first genotypic typing method applied to *C. difficile (2)*. The fact that not all *C. difficile* strains contained these extrachromosomal elements made the typeability of this method very low. In addition, strains may lose or acquire plasmids and thereby change plasmid profile *(2–4)*.

2.2. Restriction Enzyme Analysis and Restriction Fragment Length Polymorphism

Restriction enzyme, or endonuclease, analysis (REA) uses the whole genomic DNA. This DNA is digested by rare-cutter restriction enzymes, resulting in restriction fragments readable by polyacrylamide gel electrophoresis (PAGE) or agarose gel electrophoresis. The first applied REA was described by Kuijper et al. using *Hin*dIII and *Xba*I for restriction and agarose gels for analysis of the fragments *(5)*. They found that the strains detected in two patients were indistinguishable from four samples from the hospital environment, thereby showing the applicability of this method for typing *C. difficile*. They also found that the method was stable after subculturing. Another study described the use of *Cfo*I as the restriction enzyme; however, *Hin*dIII is still mostly used *(6,7)*. REA has been used as the standard typing method in North America because of its high discriminatory power and stability *(7)*, but the interpretation of REA banding patterns is

subjective, and comparative analysis of isolates has to be performed on the same gel. REA is a highly discriminatory and reproducible method; it is, however, a technically demanding procedure and very labor intensive, especially for analyzing the complex banding patterns of large numbers of isolates. For these reasons, REA data are difficult to exchange between laboratories, which is becoming an increasingly important factor for evaluating typing methods.

Restriction fragment length polymorphism (RFLP) is an alternative method that involves initial REA digestion and gel electrophoresis followed by Southern blotting with selected labeled nucleic acid probes to highlight specific restriction site heterogeneity. The difference between REA and RFLP is very small, and the designations are used interchangeably in different studies. The first description of RFLP was by Bowman et al.; restriction enzyme (*Hin*dIII) digestion was followed by gel electrophoresis and subsequent Southern blot transfer and hybridization with labeled *Escherichia coli* ribosomal RNA (rRNA) probes ***(8)***. In comparison with sodium dodecyl sulfate PAGE, immunoblotting, and REA, RFLP with an eubacterial 16S rRNA probe provided simpler patterns and yielded good discrimination ***(9)***. Another study compared the RFLP with enhanced chemiluminescence to REA, both using the *Hin*dIII enzyme for restriction ***(10)***. REA was found far more discriminatory than RFLP (34 versus 6 types among 116 isolates). REA and RFLP methods have now generally been superseded by methods based on amplification of selected targets using the polymerase chain reaction (PCR).

2.3. Arbitrary Primed PCR and Random Amplified Polymorphic DNA

Arbitrary primed PCR (AP-PCR) and random amplified polymorphic DNA (RAPD) are two methods based on PCR amplification (*see* **Chapter 4**). The primers do not have a known homology to the target sequence; subsequently, a low annealing temperature is applied. The difference between AP-PCR and RAPD is the application of a single primer versus the use of two short primers, respectively. The first described AP-PCR used six different arbitrary primers of 10–11 bp and detected six different patterns among six isolates ***(11)***.

In an outbreak among eight acquired immunodeficiency syndrome (AIDS) patients, the AP-PCR was applied using one arbitrary primer of 10 bp, differing only one nucleotide from one of the primers used by McMillin et al. ***(12)***. Among the eight isolates, seven revealed an identical AP-PCR pattern, whereas four reference strains could be discriminated from each other and the outbreak isolates. The authors concluded that the AP-PCR is simple, rapid, and discriminative for typing *C. difficile*.

Another outbreak was investigated with nearly similar arbitrary primers as in the first two studies, but a lack of reproducibility of the AP-PCR was found ***(13)***. Compared to a phenotypic method such as immunoblotting, AP-PCR resulted in better typeability

(14), and good correlation was found between AP-PCR and REA data *(15–17)*.

AP-PCR usually results in 3–12 bands between 450 and 1,300 bp, which can simply be analyzed on agarose gels. The method is cost-effective but is extremely sensitive to PCR conditions. Therefore, AP-PCR has low reproducibility, and it is difficult to establish interlaboratory comparison with this method *(18)*.

RAPD was first applied on *C. difficile* by Barbut et al. *(19)*. RAPD commonly uses two oligonucleotide primers that are short in length (ca. 10 bp) and of arbitrary sequence. Barbut et al. evaluated a RAPD method using two 10-bp primers in an investigation of CDI in AIDS patients. An identical profile was in 15 of 25 isolates, indicating a common source. RAPD compared well with pulsed-field gel electrophoresis (PFGE); while easier to perform, the results are more difficult to analyze, however *(20)*. The applicability of RAPD in the analysis of relapses versus reinfection in patients infected with the human immunodeficiency virus was shown by Alonso et al. *(21)*. Relapses were detected in 64% of patients, whereas 32% had a reinfection, and 4% had both a relapse and a reinfection *(21)*.

2.4. PCR Ribotyping

PCR ribotyping uses specific primers complimentary to sites within the RNA operons and was first applied to *C. difficile* by Gurtler, who targeted the amplification process at the spacer regions between the 16S and 23S rRNA genes *(22)*. *Clostridium difficile* was shown to possess multiple copies of the rRNA genes, which varied not only in number between strains but also in size between different copies on the same genome *(22,23)*. The method developed by Gurtler and Mayall, using radiolabeling and a long-running PAGE, detected 14 PCR ribotypes among 24 strains. The approach was simplified by Cartwright et al., who applied PCR ribotyping to 102 isolates obtained from 73 symptomatic patients *(24)*. A total of 41 types was recognized, and five of six outbreak isolates were identical *(24)*. Using the same primers as Gürtler, the PCR fragments could be separated by straightforward agarose gel electrophoresis instead of denaturing PAGE gels. The banding patterns were not affected by the quantity of DNA used in the reaction (a problem associated with AP-PCR and RAPD methods), the PCR ribotype marker was stable, and its expression was reproducible. In a comparison with the other PCR-based typing method AP-PCR, PCR ribotyping was very discriminatory and showed an agreement of 83% with PFGE, compared to 60% and 44% for AP-PCR *(25)*.

To obtain smaller fragments for better analysis on agarose gels, new primers, closer to the spacer region, were designed by O'Neill et al. in 1996 *(26)*. The amplicons, ranging from 250 to 600 bp in length, could be separated by straightforward agarose gel electrophoresis. This approach was adapted for routine use

after simplifying the method for DNA extraction *(26)*. The discriminatory power of this PCR ribotyping was compared to Delmee's serogroups, and different banding patterns were demonstrated for each of the 19 serogroups described at that time. Using these primers, at least 116 types could be discriminated within an isolate collection including nontoxinogenic and environmental strains *(27)*.

This method has since been used routinely by the U.K. Anaerobe Reference Laboratory in Cardiff, which has provided a *C. difficile* typing service for the United Kingdom since 1995. From nearly 10,000 isolates from all sources examined, a library that currently consists of over 160 distinct PCR ribotypes has been constructed. The nomenclature of types designated by this method is a three-figure numeral, and the status of this PCR ribotype library was published in 1996 when 116 types (types 001 to 116) were recognized *(27)*.

Bidet et al. further optimized PCR ribotyping using new, more specific primers based on known sequences of the 16S and 23S genes of *C. difficile (28)*. Although the method by Bidet shows better separation of bands, a large library has not yet been established as is the case with the O'Neill method, which is used worldwide *(26–29)*.

PCR ribotyping has proved a robust genotyping method, being stable and reproducible *(24,25,29–31)*. Results can be used for interlaboratory comparison and for the generation of libraries. PCR ribotyping is currently the preferred typing method in our laboratories.

2.5. Toxinotyping

Toxinotyping involves the detection of polymorphisms in the toxin A and B genes and surrounding regulatory genes, an area of the genome known collectively as the *pathogenicity locus* (PaLoc). Six regions of the PaLoc (A1–A3 and B1–B3) are amplified and digested by restriction enzymes, like in REA *(32)*. B1 and A3 are considered the most variable and are therefore good markers for detecting most toxinotypes *(33)*. Until now, 26 toxinotypes (0–X, XIa, XIb, XII–XXIV) can be discriminated among *C. difficile* strains *(32–34)*; http://www.mf.uni-mb.si/mikro/tox). Toxinotyping has been compared to serogrouping and PCR ribotyping, and a good correlation was found. Some toxinotypes are strictly associated with certain serogroups (e.g., toxinotype VIII is always seen in serogroup F strains). However, toxinotyping could further distinguish subgroups within the serogroups *(32)*. A specific PCR ribotype was usually associated with similar patterns of the toxin genes, but both methods are able to subtype each other, making toxinotyping a good addition to typing schemes *(33)*.

Barbut et al. applied the toxinotyping method on toxin A variant strains that represented 2.7% of diarrheal cases in adults

and children. Two variant types were identified by PCR of fragment A3; one type was related to toxinotype VII, due to a deletion of 600 bp in fragment A3, whereas the other type was related to toxinotype XIV, with an insertion of about 200 bp *(35)*. In a study of 153 clinical isolates from an American hospital, 11.1% of strains belonged to toxinotypes other than toxinotype 0. An additional toxin, the binary toxin, was found only in nine strains, all of which were variant toxinotypes *(36)*.

The reproducibility of the method is 100%, and the discriminatory power is good, although, for example, PFGE and PCR ribotyping show more discrimination between strains. The most important advantage of this typing method is that a clear view of the toxin status of *C. difficile* strains can be acquired.

3. Recently Developed Typing Methods for *C. difficile*

3.1. flic*C* Typing

An alternative PCR target for typing purposes is the flagellin gene *flicC*, described by Tasteyre et al. *(37)*. In a study of 47 isolates belonging to 11 different serogroups, three profiles could be recognized. When the method was expanded with RFLP analysis, nine different RFLP patterns were recognized. Although nonflagellated strains were included, they did contain the *fli*C gene. In a study with nine toxin A–/B+ strains, only three strains showed flagella. However, all nine strains belonged to the same type using *fliC* PCR-RFLP *(38)*.

3.2. slp*A* Typing

Another gene studied for typing is the *slp*A gene, encoding an S-layer precursor protein of *C. difficile*. Seven S types have been recognized, of which one type accounted for 73% of the clinical cases and 93% of the environmental cases *(39)*. Thirty-two strains belonging to ten serogroups were used for PCR-RFLP and sequencing analysis of the variable region. This RFLP-sequence combination led to sequences identical within a given serogroup and differences between serogroups and was therefore thought of as an alternative typing method for *C. difficile (40)*. The *slpA* genotyping by PCR-RFLP was subsequently tested on Japanese outbreak strains and resulted in three subtypes. The method was also applied directly on fecal samples, and results were in complete agreement with the cultured strains from these samples *(41)*. Typing of *slp*A is considered a reproducible method with the advantage of interlaboratory data exchange. The s*lp*A typing of strains of 14 different PCR ribotypes identified 9 groups; PCR ribotypes showed completely identical *slp*A sequence in two cases and 1- to 3-bp differences within other groups *(42)*.

3.3. Amplified Fragment Length Polymorphism

Amplified fragment length polymorphism (AFLP) has also been applied as a typing method for *C. difficile* ***(43)***. The AFLP method uses restriction, ligation, and selective amplification on the whole genome. Differentiation can be made due to variation per type in restriction site mutations, mutations in the sequences adjacent to the restriction sites and complementary to the selective primer extensions, and insertions and deletions within the amplified fragments. While the reproducibility of AFLP was similar to PFGE, Klaassen et al. showed that the typeability of AFLP was better, especially for isolates for which some DNA degradation had occurred. In addition, AFLP was found to be faster and easier to perform on small quantities of DNA ***(43)***. Analysis of 30 clinical isolates encompassing all known sero(sub)groups and of 30 PCR ribotype 017 toxin A–/B+ isolates from various countries showed that the discriminatory power of AFLP was similar to that of PFGE ***(44)***.

3.4. Multilocus Sequence Typing

Multilocus sequence typing (MLST) has also been tested as a typing method for *C. difficile*. MLST consists of DNA sequence analysis of housekeeping genes after PCR amplification and is mostly used to study genetic relationships and population structures ***(45)***. MLST developed for *C. difficile* includes seven housekeeping genes. Among 72 isolates from various origins, 62 PCR ribotypes and 34 sequence types (STs) could be discriminated. In a dendrogram representing the relationships between the STs, three divergent lineages could be recognized, of which one strictly contained toxin A–/B+ strains ***(45)***. The method was further expanded by the inclusion of ten virulence-associated genes, among which were *fli*C, *slp*A, *tcd*A, *tcd*B, and *tcd*D ***(46)***. A total of 29 isolates from various origins and representing 22 STs selected from the lineages found in their first study were investigated. The polymorphisms detected in the virulence-associated genes were comparable to those of the housekeeping genes. However, *cwp66* and *slpA* appeared highly polymorphic, although only 11 and 16 alleles could be detected, respectively. Again, toxin A–/B+ strains belonged to a homogeneous lineage, and a fourth lineage could be characterized in contrast to the method based on only housekeeping genes ***(46)***. No association was found between the STs and the clinical presentation or the source of the isolates ***(45,46)***. It was concluded that MLST with the virulence-associated genes included was more discriminatory than the housekeeping genes alone, although this could depend on the genes chosen. The main advantage of the method is the yield of unambiguous sequence data. No comparisons with other techniques have been described to date.

3.5. Multilocus Variable Number of Tandem Repeats Analysis

The analysis of the sequenced human and bacterial genomes revealed a high percentage of DNA that consisted of a variable number of tandem repeats (VNTR). The repeats vary in size,

location, complexity, and repeat mode and can occur clustered in one genomic area or dispersed throughout the entire genome. These repeat arrays can be targets for genomic events, such as DNA polymerase slippage and recombination. It is the polymorphic property of the VNTRs that led to the application in identification and typing of bacteria. Multilocus VNTR analysis (MLVA) has already been tested successfully on a number of bacterial species due to its high reproducibility, high discriminatory power, and typeability ***(47)***. The availability of the complete sequence of the *C. difficile* genome of strain 630 (http://www.sanger.ac.uk/Projects/C_difficile/;***48)*** provided the opportunity to identify these short tandem repeats.

The MLVA developed by Marsh et al. uses automated sequence detection and subsequent manual determination of the number of tandem repeats per locus ***(49)***. Seven short tandem repeat loci were amplified from 40 isolates from two different sources, and REA was tested on every strain as well. The stability was good, although differences of one repeat could arise. This MLVA clustered outbreak strains of the same REA type and discriminated different REA types from each other.

For a faster and easier application of the MLVA for *C. difficile*, a new method was developed using smaller short tandem repeats (2–9 bp) to facilitate automated fragment analysis with multicolored capillary electrophoresis instead of sequencing ***(50)***. This MLVA technique was compared to PCR ribotyping and tested on a set of 56 reference strains encompassing 31 serogroups and 25 toxinotypes. In addition, clinical isolates were included from outbreaks in different countries due to the new emerging type 027 and the toxin A–/B+ strain PCR ribotype 017. Of seven VNTR, four were identical to those used in the study of Marsh et al. ***(49)***. MLVA was highly (100%) reproducible with an excellent stability of all seven loci. All tested PCR ribotypes could be recognized, including the seven subtypes of 001. In contrast to PCR ribotyping, MLVA was able to discriminate strains belonging to serogroups A7 from A11, A9 from A10, A8 from S1, H from K, and A14 from S4. Toxin A–/B+ strains could be recognized as eight country-specific clusters. All strains with 100% similarity belonged to country-specific clusters. Interestingly, toxin A–/B+ strain could be differentiated from all other types using the combination of two markers. PCR ribotype 027 strains from several outbreaks in the Netherlands were clustered in 14 different groups using MLVA; the clusters were mostly hospital specific. All strains were completely identical to each other with the combination of three markers with 10, 4, and 2 repeats, respectively. Only the U.K. strain showed six repeats for a marker instead of ten, indicating a possible difference between type 027 strains from specific countries.

The utility of MLVA and PFGE to identify clusters of CDI was tested among 91 isolates of PCR ribotype 027 (NAP1, for North

American pulsed-field type 1) from nine hospitals in England ***(51)***. PFGE discriminated between ribotype 027 strains at greater than 98% similarity, identifying five pulsovars (I to V) with 1 to 53 isolates each. MLVA was markedly more discriminatory, identifying 23 types with 1 to 15 isolates (>71% similarity). MLVA was far superior to PFGE for analyzing clusters of CDI both within and between institutions.

In a study using isolates from laboratories in Canada, the Netherlands, the United Kingdom, and the United States, seven *C. difficile* typing techniques were compared: MLVA, AFLP, *slpA* sequence typing, PCR ribotyping, REA, MLST, and PFGE ***(52)***. All 42 isolates were typeable by all techniques, but only REA and MLVA showed sufficient discrimination to distinguish strains from different outbreaks ***(52)***. MLVA has also been applied to study local outbreaks of clindamycin-resistant *C. difficile* PCR ribotype 027 strains ***(53,54)***.

4. Conclusions

All typing methods have certain advantages and disadvantages, but their ultimate contribution to knowledge is dictated by their performance according to the criteria listed by Struelens: typeability, reproducibility, stability, discriminatory power, and epidemiological concordance. It should also have technical advantages, such as ease of performance, relative low cost, and high throughput. In due course, as new methods come and go, one method will probably emerge as the most suitable. Currently, the PCR-ribotyping method and library of PCR ribotypes in Cardiff are the benchmark to which most typing studies around the world are compared, and more important, they have probably contributed most to our current knowledge of the global epidemiology of *C. difficile*. Undoubtedly, further advances in molecular subtyping methods will add even further to our understanding of the epidemiology this important nosocomial pathogen.

References

1. Kuijper, E. J., Coignard, B., and Tull, P. (2006). Emergence of *Clostridium difficile*-associated disease in North America and Europe. *Clin. Microbiol. Infect.* **12**(Suppl. 6), 2–18.
2. Arai, T., Kusakabe, A., Nakashio, S., and Nakamura, M. (1984). A survey of plasmids in *Clostridium difficile* strains. *Kitasato Arch. Exp. Med.* **57**, 285–288.
3. Steinberg, J. P., Beckerdite, M. E., and Westenfelder, G. O. (1987). Plasmid profiles of *Clostridium difficile* isolates from patients with antibiotic-associated colitis in two community hospitals. *J. Infect. Dis.* **156**, 1036–1038.
4. Clabots, C., Lee, S., Gerding, D., Mulligan, M., Kwok, R., Schaberg, D., et al. (1988). *Clostridium difficile* plasmid isolation as an

epidemiologic tool. *Eur. J. Clin. Microbiol. Infect. Dis.* **7**, 312–315.

5. Kuijper, E. J., Oudbier, J. H., Stuifbergen, W. N., Jansz, A., and Zanen, H. C. (1987). Application of whole-cell DNA restriction endonuclease profiles to the epidemiology of *Clostridium difficile*-induced diarrhea. *J. Clin. Microbiol.* **25**, 751–753.
6. Devlin, H. R., Au, W., Foux, L., and Bradbury, W. C. (1987). Restriction endonuclease analysis of nosocomial isolates of *Clostridium difficile*. *J. Clin. Microbiol.* **25**, 2168–2172.
7. Clabots, C. R., Johnson, S., Bettin, K. M., Mathie, P. A., Mulligan, M. E., Schaberg, D. R., et al. (1993). Development of a rapid and efficient restriction endonuclease analysis typing system for *Clostridium difficile* and correlation with other typing systems. *J. Clin. Microbiol.* **31**, 1870–1875.
8. Bowman, R. A., O'Neill, G. L., and Riley, T. V. (1991). Non-radioactive restriction fragment length polymorphism (RFLP) typing of *Clostridium difficile*. *FEMS Microbiol. Lett.* **63**, 269–272.
9. Wolfhagen, M. J., Fluit, A. C., Torensma, R., Jansze, M., Kuypers, A. F., Verhage, E. A., et al. (1993). Comparison of typing methods for *Clostridium difficile* isolates. *J. Clin. Microbiol.* **31**, 2208–2211.
10. O'Neill, G. L., Beaman, M. H., and Riley, T. V. (1991). Relapse versus reinfection with *Clostridium difficile*. *Epidemiol. Infect.* **107**, 627–635.
11. McMillin, D. E., and Muldrow, L. L. (1992). Typing of toxic strains of *Clostridium difficile* using DNA fingerprints generated with arbitrary polymerase chain reaction primers. *FEMS Microbiol. Lett.* **71**, 5–9.
12. Barbut, F., Mario, N., Frottier, J., and Petit, J. C. (1993). Use of the arbitrary primer polymerase chain reaction for investigating an outbreak of *Clostridium difficile*-associated diarrhea in AIDS patients. *Eur. J. Clin. Microbiol. Infect. Dis.* **12**, 794–795.
13. Wilks, M., and Tabaqchali, S. (1994). Typing of *Clostridium difficile* by polymerase chain reaction with an arbitrary primer. *J. Hosp. Infect.* **28**, 231–234.
14. Killgore, G. E., and Kato, H. (1994). Use of arbitrary primer PCR to type *Clostridium difficile* and comparison of results with those by immunoblot typing. *J. Clin. Microbiol.* **32**, 1591–1593.
15. Tang, Y. J., Houston, S. T., Gumerlock, P. H., Mulligan, M. E., Gerding, D. N., Johnson, S., et al. (1995). Comparison of arbitrarily primed PCR with restriction endonuclease and immunoblot analyses for typing *Clostridium difficile* isolates. *J. Clin. Microbiol.* **33**, 3169–3173.
16. Samore, M., Killgore, G., Johnson, S., Goodman, R., Shim, J., Venkataraman, L., et al. (1997). Multicenter typing comparison of sporadic and outbreak *Clostridium difficile* isolates from geographically diverse hospitals. *J. Infect. Dis.* **176**, 1233–1238.
17. Rafferty, M. E., Baltch, A. L., Smith, R. P., Bopp, L. H., Rheal, C., Tenover, F. C., et al. (1998). Comparison of restriction enzyme analysis, arbitrarily primed PCR, and protein profile analysis typing for epidemiologic investigation of an ongoing *Clostridium difficile* outbreak. *J. Clin. Microbiol.* **36**, 2957–2963.
18. Cohen, S. H., Tang, Y. J., and Silva, J., Jr. (2001). Molecular typing methods for the epidemiological identification of *Clostridium difficile* strains. *Expert. Rev. Mol. Diagn.* **1**, 61–70.
19. Barbut, F., Mario, N., Delmee, M., Gozian, J., and Petit, J. C. (1993). Genomic fingerprinting of *Clostridium difficile* isolates by using a random amplified polymorphic DNA (RAPD) assay. *FEMS Microbiol. Lett.* **114**, 161–166.
20. Chachaty, E., Saulnier, P., Martin, A., Mario, N., and Andremont, A. (1994). Comparison of ribotyping, pulsed-field gel electrophoresis and random amplified polymorphic DNA for typing *Clostridium difficile* strains. *FEMS Microbiol. Lett.* **122**, 61–68.
21. Alonso, R., Gros, S., Pelaez, T., Garcia-de-Viedma, D., Rodriguez-Creixems, M., and Bouza, E. (2001). Molecular analysis of relapse vs re-infection in HIV-positive patients suffering from recurrent *Clostridium difficile* associated diarrhoea. *J. Hosp. Infect.* **48**, 86–92.
22. Gurtler, V. (1993). Typing of *Clostridium difficile* strains by PCR-amplification of variable length 16S-23S rDNA spacer regions. *J. Gen. Microbiol.* **139**, 3089–3097.
23. Gurtler, V., and Mayall, B. C. (1994). Genotyping of *Clostridium difficile* isolates. *J. Clin. Microbiol.* **32**, 3095.
24. Cartwright, C. P., Stock, F., Beekmann, S. E., Williams, E. C., and Gill, V. J. (1995). PCR amplification of rRNA intergenic spacer regions as a method for epidemiologic typing of *Clostridium difficile*. *J. Clin. Microbiol.* **33**, 184–187.
25. Collier, M. C., Stock, F., DeGirolami, P. C., Samore, M. H., and Cartwright, C. P. (1996). Comparison of PCR-based approaches to molecular epidemiologic analysis of *Clostridium difficile*. *J. Clin. Microbiol.* **34**, 1153–1157.
26. O'Neill, G. L., Ogunsola, F. T., Brazier, J. S., and Duerden, B. I. (1996). Modification of a PCR ribotyping method for application as a

routine typing scheme for *Clostridium difficile*. *Anaerobe* **2**, 205–209.

27. Stubbs, S. L., Brazier, J. S., O'Neill, G. L., and Duerden, B. I. (1999). PCR targeted to the 16S-23S rRNA gene intergenic spacer region of *Clostridium difficile* and construction of a library consisting of 116 different PCR ribotypes. *J. Clin. Microbiol.* **37**, 461–463.
28. Bidet, P., Barbut, F., Lalande, V., Burghoffer, B., and Petit, J. C. (1999). Development of a new PCR ribotyping method for *Clostridium difficile* based on ribosomal RNA gene sequencing. *FEMS Microbiol. Lett.* **175**, 261–266.
29. Barbut, F., Richard, A., Hamadi, K., Chomette, V., Burghoffer, B., and Petit, J. C. (2000). Epidemiology of recurrences or reinfections of *Clostridium difficile*-associated diarrhea. *J. Clin. Microbiol.* **38**, 2386–2388.
30. Brazier, J. S., Mulligan, M. E., Delmee, M., and Tabaqchali, S. (1997). Preliminary findings of the international typing study on *Clostridium difficile*. International Clostridium Difficile Study Group. *Clin. Infect. Dis.* **25**(Suppl. 2), S199–S201.
31. Brazier, J. S. (2001). Typing of *Clostridium difficile*. *Clin. Microbiol. Infect.* **7**, 428–431.
32. Rupnik, M., Avesani, V., Janc, M., von Eichel-Streiber, C., and Delmee, M. (1998). A novel toxinotyping scheme and correlation of toxinotypes with serogroups of *Clostridium difficile* isolates. *J. Clin. Microbiol.* **36**, 2240–2247.
33. Rupnik, M., Brazier, J. S., Duerden, B. I., Grabnar, M., and Stubbs, S. L. (2001). Comparison of toxinotyping and PCR ribotyping of *Clostridium difficile* strains and description of novel toxinotypes. *Microbiology* **147**, 439–447.
34. Rupnik, M., Kato, N., Grabnar, M., and Kato, H. (2003). New types of toxin A-negative, toxin B-positive strains among *Clostridium difficile* isolates from Asia. *J. Clin. Microbiol.* **41**, 1118–1125.
35. Barbut, F., Lalande, V., Burghoffer, B., Thien, H. V., Grimprel, E., and Petit, J. C. (2002). Prevalence and genetic characterization of toxin A variant strains of *Clostridium difficile* among adults and children with diarrhea in France. *J. Clin. Microbiol.* **40**, 2079–2083.
36. Geric, B., Rupnik, M., Gerding, D. N., Grabnar, M., and Johnson, S. (2004). Distribution of *Clostridium difficile* variant toxinotypes and strains with binary toxin genes among clinical isolates in an American hospital. *J. Med. Microbiol.* **53**, 887–894.
37. Tasteyre, A., Karjalainen, T., Avesani, V., Delmee, M., Collignon, A., Bourlioux, P., et al. (2000). Phenotypic and genotypic diversity of the flagellin gene (*fli*C) among *Clostridium difficile* isolates from different serogroups. *J. Clin. Microbiol.* **38**, 3179–3186.
38. Pituch, H., Obuch-Woszczatynski, P., van den Braak, N., van Belkum, A., Kujawa, M., Luczak, M., et al. (2002). Variable flagella expression among clonal toxin A–/B+ *Clostridium difficile* strains with highly homogeneous flagellin genes. *Clin. Microbiol. Infect.* **8**, 187–188.
39. McCoubrey, J., Starr, J., Martin, H., and Poxton, I. R. (2003). *Clostridium difficile* in a geriatric unit: a prospective epidemiological study employing a novel S-layer typing method. *J. Med. Microbiol.* **52**, 573–578.
40. Karjalainen, T., Saumier, N., Barc, M. C., Delmee, M., and Collignon, A. (2002). *Clostridium difficile* genotyping based on *slp*A variable region in S-layer gene sequence: an alternative to serotyping. *J. Clin. Microbiol.* **40**, 2452–2458.
41. Kato, H., Yokoyama, T., and Arakawa, Y. (2005). Typing by sequencing the *slp*A gene of *Clostridium difficile* strains causing multiple outbreaks in Japan. *J. Med. Microbiol.* **54**, 167–171.
42. Eidhin, D. N., Ryan, A. W., Doyle, R. M., Walsh, J. B., and Kelleher, D. (2006). Sequence and phylogenetic analysis of the gene for surface layer protein, *slp*A, from 14 PCR ribotypes of *Clostridium difficile*. *J. Med. Microbiol.* **55**, 69–83.
43. Klaassen, C. H., van Haren, H. A., and Horrevorts, A. M. (2002). Molecular fingerprinting of *Clostridium difficile* isolates: pulsed-field gel electrophoresis versus amplified fragment length polymorphism. *J. Clin. Microbiol.* **40**, 101–104.
44. van den Berg, R. J., Claas, E. C., Oyib, D. H., Klaassen, C. H., Dijkshoorn, L., Brazier, J. S., et al. (2004). Characterization of toxin A-negative, toxin B-positive *Clostridium difficile* isolates from outbreaks in different countries by amplified fragment length polymorphism and PCR ribotyping. *J. Clin. Microbiol.* **42**, 1035–1041.
45. Lemee, L., Dhalluin, A., Pestel-Caron, M., Lemeland, J. F., and Pons, J. L. (2004). Multilocus sequence typing analysis of human and animal *Clostridium difficile* isolates of various toxigenic types. *J. Clin. Microbiol.* **42**, 2609–2617.
46. Lemee, L., Bourgeois, I., Ruffin, E., Collignon, A., Lemeland, J. F., and Pons, J. L. (2005). Multilocus sequence analysis and comparative evolution of virulence-associated genes and housekeeping genes of *Clostridium difficile*. *Microbiology* **151**, 3171–3180.
47. Lindstedt, B.-A. (2005). Multiple-locus variable number tandem repeats analysis for genetic fingerprinting of pathogenic bacteria. *Electrophoresis* **26**, 2567–2582.

48. Sebaihia, M., Wren, B. W., Mullany, P., Fairweather, N. F., Minton, N., Stabler, R., et al. (2006). The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. *Nat. Genet.* **38**, 779–786.
49. Marsh, J. W., O'Leary, M. M., Shutt, K. A., Pasculle, A. W., Johnson, S., Gerding, D. N., et al. (2006). Multilocus variable-number tandem-repeat analysis for investigation of *Clostridium difficile* transmission in Hospitals. *J. Clin. Microbiol.* **44**, 2558–2566.
50. van den Berg, R. J., Schaap, I., Templeton, K. E., Klaassen, C. H., and Kuijper, E. J. (2007). Typing and subtyping of *Clostridium difficile* isolates by using multiple-locus variable-number tandem-repeat analysis. *J. Clin. Microbiol.* **45**, 1024–1028.
51. Fawley, W. N., Freeman, J., Smith, C., Harmanus, C., van den Berg, R. J., Kuijper, E. J., et al. (2008). Use of highly discriminatory fingerprinting to analyze clusters of *Clostridium difficile* infection cases due to epidemic ribotype 027 strains. *J. Clin. Microbiol.* **46**, 954–960.
52. Killgore, G., Thompson, A., Johnson, S., Brazier, J., Kuijper, E., Pepin, J., et al. (2008). Comparison of seven techniques for typing international epidemic strains of *Clostridium difficile*: restriction endonuclease analysis, pulsed-field gel electrophoresis, PCR-ribotyping, multilocus sequence typing, multilocus variable-number tandem-repeat analysis, amplified fragment length polymorphism, and surface layer protein A gene sequence typing. *J. Clin. Microbiol.* **46**, 431–437.
53. Fenner, L., Widmer, A. F., Stranden, A., Conzelmann, M., Goorhuis, A., Harmanus, C., et al. (2008). First cluster of clindamycin-resistant *Clostridium difficile* PCR-ribotype 027 in Switzerland. *Clin. Microbiol.* **14**, 514–515.
54. Drudy, D., Goorhuis, B., Bakker, D., Kyne, L., van den Berg, R., Fanning, S., et al. (2008). Clindamycin-resistant clone of *Clostridium difficile* PCR ribotype 027, Europe. *Emerg. Infect. Dis.* **14**, 1485–1487.

Chapter 14

Genotyping of *Mycobacterium tuberculosis* Clinical Isolates Using IS*6110*-Based Restriction Fragment Length Polymorphism Analysis

Pablo Bifani, Natalia Kurepina, Barun Mathema, Xiao-Ming Wang, and Barry Kreiswirth

Abstract

A number of phylogenetic studies of *Mycobacterium tuberculosis* have suggested a highly clonal population structure. Despite the extreme homogeneity of *M. tuberculosis* strains, the genome is punctuated by a number of polymorphic regions that give rise to sufficient diversity, thus forming the basis for molecular epidemiologic studies of tuberculosis. As such, insertion sequence (IS) *6110*, which is unique to members of the *M. tuberculosis* complex and is present in variable numbers and in discrete genomic locales among strains, has been extensively used in molecular epidemiologic studies. Genotyping, using IS*6110*-based restriction fragment length polymorphism (RFLP), was standardized by the international community, and this has facilitated inter- and intralaboratory comparison, thereby serving as a model system for subspeciation of *M. tuberculosis*. When IS*6110*-based RFLP was used in conjunction with conventional epidemiologic data, its utility was realized. In this chapter, we discuss the basic methodology for conducting IS*6110*-based RFLP and analyzing the resulting hybridization profiles.

Key words: IS*6110* insertion sequence, molecular epidemiology, *Mycobacterium tuberculosis*, Southern blot hybridization.

1. Introduction

Tuberculosis (TB) remains one of the world's most prevalent infectious diseases, accounting for 9 million new cases and 1.7 million deaths in 2006 alone *(1)*. The problem is exacerbated by the growing number of individuals with TB coinfected with the human immunodeficiency virus, as well as by cases of multidrug-resistant

D.A. Caugant (ed.), *Molecular Epidemiology of Microorganisms*, Methods in Molecular Biology, Vol. 551
DOI: 10.1007/978-1-60327-999-4_14,

TB (MDR-TB) and extensively drug-resistant TB (XDR-TB). A report by the World Health Organization estimated 458,000 individuals with MDR-TB globally *(2)*, underlining the urgent need to control this disease. Although efforts are being made in search of new antimycobacterial compounds, various therapies, and vaccine development, history has proven that the most efficacious means of controlling infections relies on the implementation of adequate public health measures and improvement of basic living conditions.

Thus, to improve TB control efforts, a more thorough understanding of TB epidemiology is essential. This can be achieved by elucidating transmission dynamics, the contribution of recent versus reactive disease, as well as the nature and extent of drug resistance among studied populations. Due to the intrinsic characteristics of TB's etiologic agent, *Mycobacterium tuberculosis*, that is, slow growth (~24 h doubling time), a long latency period, and airborne route of infection, some key epidemiologic questions remain elusive. Recently, however, our understanding of TB epidemiology has benefited extensively from the integration of molecular techniques with conventional epidemiologic data; which is known as *molecular epidemiology*. This is providing greater resolution to address previously unanswered questions relevant to TB control and prevention. With increased access to genomic information, a number of techniques have been developed to genotype or fingerprint *M. tuberculosis*. In some cases, these techniques have been implemented on a large number of clinical isolates from diverse geographic and epidemiologic sources, thereby rendering *M. tuberculosis* genotyping a model system in the nascent field of molecular epidemiology. Here we discuss some of the characteristics and applications of insertion sequence (IS) *6110*-based fingerprinting of *M. tuberculosis* isolates for molecular epidemiological studies. The specific protocols for *M. tuberculosis* strain genotyping may vary according to the resources available and on approval by the institutional biosafety committees of the local or national laboratories.

1.1. Characteristics of the M. tuberculosis *IS*6110

Mycobacterium tuberculosis is a member of group of closely related species, collectively known as the *M. tuberculosis* complex (MTBC), which is comprised of seven members (*M. tuberculosis* sensu stricto, *Mycobacterium africanum, Mycobacterium bovis, Mycobacterium canetti, Mycobacterium microti, Mycobacterium caprae,* and *Mycobacterium pinnipedii*). Among other unique features, such as extreme genetic homogeneity and wide host-specific ranges, the MTBC bears a unique IS, IS*6110*. IS*6110* is a transposable element that is a member of the IS3 family of ISs or mobile elements (*3*).

Briefly, IS*6110* is a 1,355-bp long fragment that encodes four enzymes required for its own transposition and insertion and is

flanked by imperfect 28-bp inverted repeats *(4)*. On insertion, a 3- to 4-bp target duplication is generated, and loss is usually accompanied by deletions (genomic scars) of the flanking regions. IS*6110* may transpose into functional genes or regulatory sequences and hence can alter gene expression and subsequently the protein profile, thus in some cases altering the phenotype *(5)*. IS*6110* is found in virtually all *M. tuberculosis* strains (**Fig. 1**), with some bearing up to 24 copies, although there exist strains, albeit rarely, with no IS*6110* elements. The factors that drive IS*6110* transposition have not yet been fully understood, but transposition has been shown not to be dependent on sequence variation of the element itself as IS*6110* is highly conserved throughout (*6*).

The observed diversity in IS*6110* copy number and genomic position between unrelated strains of *M. tuberculosis* can be utilized to examine microevolutionary processes. In addition, phylogenetic studies using synonymous single-nucleotide polymorphisms

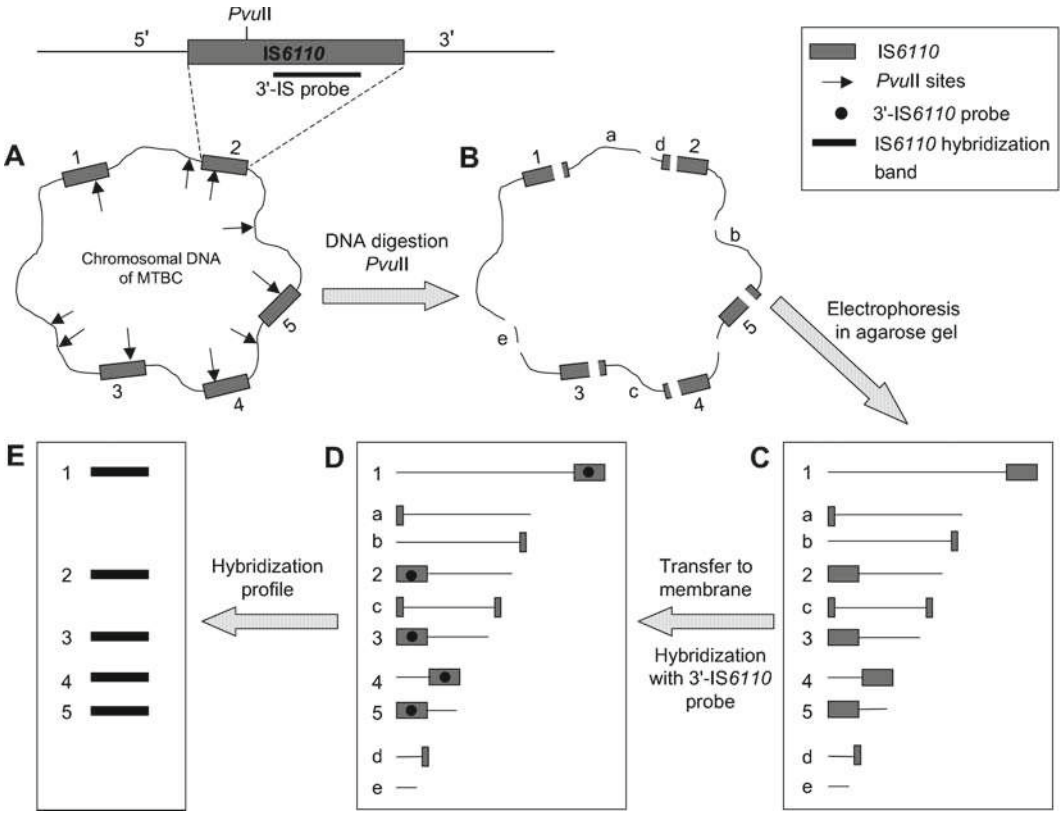


Fig. 1. The schema of the principal procedures in IS*6110*-based RFLP genotyping. (**A**) Localization of five IS*6110* copies in the *M. tuberculosis* genome (positions and orientation of IS may be different in different clinical isolates); (**B**) *Pvu*II digest of chromosomal DNA; (**C**) electrophoresis in agarose gel distributes DNA fragments according to their molecular weight (size); (**D**) the presence of IS*6110* 3′-specific arm in chromosomal DNA fragments; (**E**) bands revealed after hybridization with IS*6110*-specific probe.

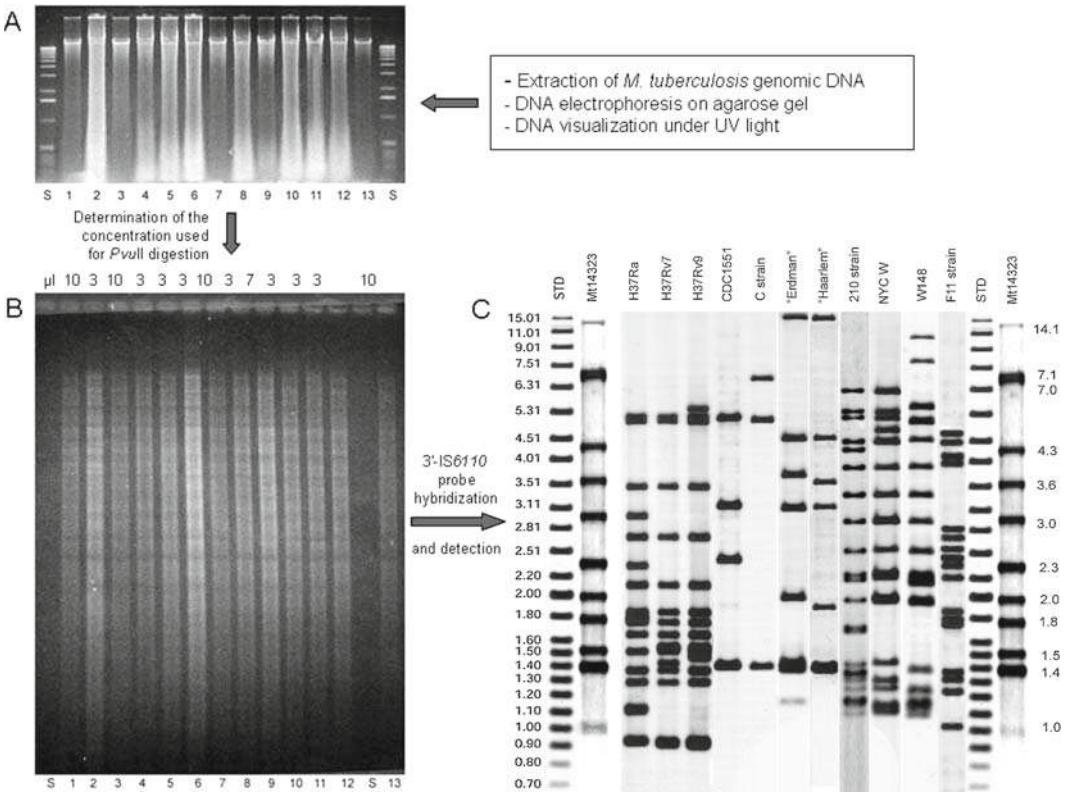


Fig. 2. Example of hybridization blot in different stages of DNA analysis. (**A**) Electrophoresis of total chromosomal DNA, 5 µL of each sample loaded. Calculation of the amount of chromosomal DNA needed for hybridization (bottom: sample number; under: volume of DNA required). (**B**) Electrophoresis of *Pvu*II-digested chromosomal DNA, stained with ethidium bromide. (**C**) Blot after overnight hybridization with IS*6110*-specific probe and detection procedures.

(sSNPs) have shown that IS*6110* copy number and location are similar within discrete lineages (or clades), underlying evolutionary significance, as illustrated in **Fig. 2**. Not surprisingly, isolates with distinct IS*6110* profiles seem to aggregate socially and therefore track to geographic locales where they may be endemic.

*1.2. IS*6110*-Based Restriction Fragment Length Polymorphisms*

One key aspect underlying the success of molecular epidemiologic studies of *M. tuberculosis* is the implementation and adoption of a standardized protocol for IS*6110*-Southern blot hybridization. This standardized protocol allowed for inter- and intralaboratory comparative analysis and set a precedent for other bacterial genotyping systems. The standardization involved (i) the use of restriction endonuclease *Pvu*II for *M. tuberculosis* genomic DNA digestion, yielding a smear of *Pvu*II-flanked fragments; (ii) the selection of the 3′ (right-side) fragment of IS*6110* as a

hybridization probe; and (iii) a standardized molecular weight marker and technical recommendations for conducting IS*6110*-based restriction fragment length polymorphism (RFLP) *(7)*. Briefly, as shown in **Fig. 3**, *Pvu*II-restricted fragments of chromosomal DNA are separated by electrophoresis, blotted onto a nitrocellulose membrane and hybridized with the 3′-IS*6110* fragment as a hybridization probe (the IS*6110* elements contain one asymmetrical *Pvu*II restriction site with 3′-arm 900-bp long). The DNA fragment serving as the IS*6110*-specific probe can be generated by polymerase chain reaction (PCR) using *M. tuberculosis* chromosomal DNA. Alternatively, an *Escherichia coli* plasmid, such as pUC18, containing the cloned 3′-IS*6110* fragment as the target DNA for amplification can be used. This plasmid containing a 312-bp 3′-IS*6110* fragment (pUCIS) can be requested from Dr. Kurepina at PHRI (Newark, NJ). The probe can be labeled using radioactive ^{32}P or with a chemiluminescence kit (ECL, GE Healthcare, UK; GE Healthcare Life Sciences, Piscataway, NJ). Following hybridization and detection, the number of bands (= number of IS*6110* copies per genome) and patterns of bands on

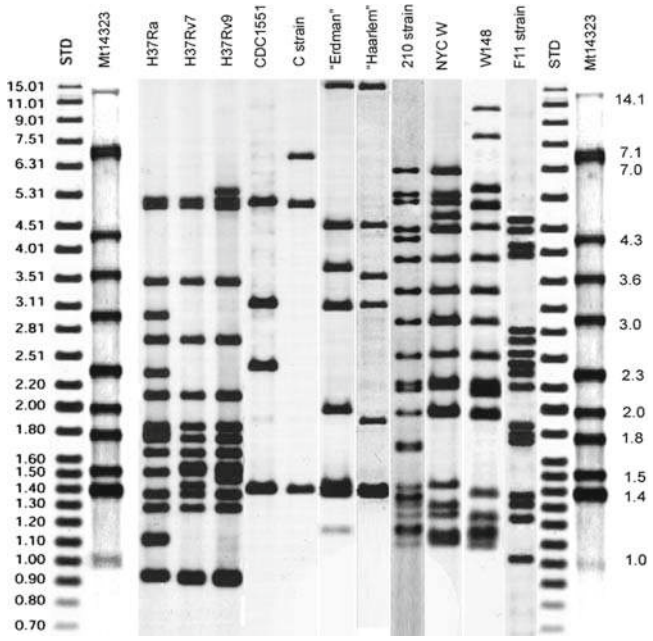


Fig. 3. IS*6110*-based RFLP images of the most known and annotated clinical and laboratory *M. tuberculosis* strains. STD, standards, developed by the U.S. Centers for Disease Control and Prevention (CDC); Mt14323, strain used as standard for *M. tuberculosis* comparison. H37Ra, H37Rv7, and H37Rv9 are laboratory strains; CDC1551, C strain, "Erdman" and "Haarlem" (strains completely sequenced); W strain, W148, and F11 (sequencing in progress).

the blot (size of *Pvu*II-flanked chromosomal fragment containing the right-side IS*6110* band) form the fingerprint of the strain.

1.3. Applications of IS6110 RFLP in TB Epidemiology

The main objective of molecular epidemiology is to associate particular epidemiologic data with strain relatedness, for example, to identify chains of transmission. In general, IS*6110* RFLP analysis has been a highly valuable tool when analyzing populations of *M. tuberculosis* isolates. The stability of IS*6110* has proven to be sufficient to identify the same strains spreading from one patient to another, thus, implicating transmission. Yet, IS*6110* is sufficiently mobile to show diversity within a given population. The two criteria, stability and diversification over the time (biological clock), are fulfilled by IS*6110,* making it a suitable epidemiological marker. The exception lies in *M. tuberculosis* isolates comprising fewer than six copies of IS*6110*; thus, by convention, it has been agreed that RFLP pattern discrimination is reliable for samples possessing six or more IS*6110* copies. This limitation does not represent a problem in cosmopolitan cities, where strain diversity is extensive. However, in some geographical regions, such as southern India, isolates carrying one or two insertions are predominant, hence rendering this analytical tool futile.

IS*6110* RFLP has proven to be highly significant in the analysis of samples harboring multiple IS*6110* copy number isolates. For example, the W-Beijing strains (with more than eight IS*6110* elements) that are predominant in China, Southeast Asia, and Eurasia can be sufficiently discriminated by IS*6110* RFLP analysis. The relatedness and frequency of certain strain types in the populations can be determined, as is the case with other groups of related strains (e.g., Harleem strains).

1.4. Combining IS6110 with Other Molecular Techniques for Specific Strain Identification

In general, it can be assumed that isolates with more than five IS*6110* copies displaying different IS*6110* RFLP profiles may be cases of reactive disease, while clusters (i.e., identical IS*6110* profiles) may represent a transmission event. Isolates with fewer than six bands need to be analyzed by other unrelated molecular techniques to draw proper inference on chains of transmission. Other molecular markers can be spoligotyping, mycobacterial interspersed repetitive unit variable number tandem repeat (MIRU-VNTR) analysis, deligotyping, SNPs, or any particular characteristics of the isolates in question, including unique sequences, duplications, deletions, insertions, or drug resistance phenotypes *(8)*. In addition, coupling IS*6110* fingerprinting with other molecular markers can be useful in identifying clonal strains and substrain families or following the microevolution of a given strain in a population. Such specific markers have been used to determine the clonal nature of the W-Beijing strain family *(9)*, to identify substrain families of epidemiologically related W-Beijing phenotypes *(10–12)*, and to follow the microevolution of isolates with

few IS*6110* elements *(13)*. Deletion of spacers within the direct repeat locus (*see* **Chapter 10**) can be in some cases associated with an IS*6110* band shift, as shown in **Fig. 4** for strains BW90 and BW900 or strain W14 (*12*).

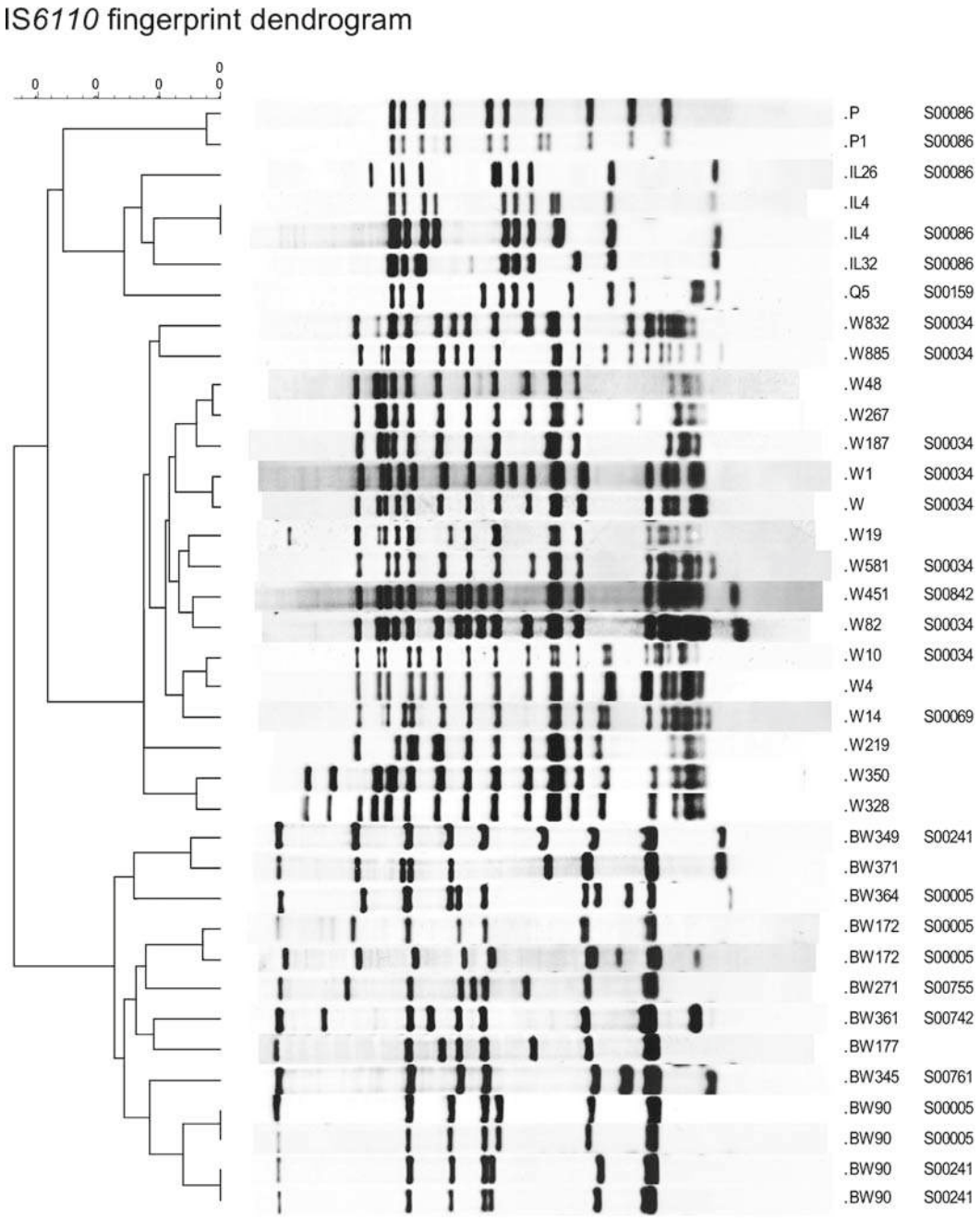


Fig. 4. Dendrogram of clinical *M. tuberculosis* strains (Bionumerics, Applied Math, Belgium). These three groups of isolates correspond to the spoligofamilies identified by spoligotyping technique (*see* **Chapter 10**). Strains W and W1 differ by one band and were part of the same outbreak in New York City in the 1990s.

2. Materials

2.1. Media for *M. tuberculosis* Culture

1. Preferred choice is Lowenstein Jensen (LJ) (Difco or homemade) slants.
2. Alternatively, use Middlebrook 7H11 agar (Difco). Dissolve 21 g in 900 mL deionized water, supplement with 10 mL of 50% glycerol (sterile stock 50/50 glycerol/deionized water). Melt the agar at 100°C and autoclave at 121°C for 15 min. Cool to 55°C and add 100 mL of prewarmed (37°C) Middlebrook OADC supplement (Difco). Mix and pour plates.

2.2. DNA Isolation

1. Proteinase K (Boehringer Mannheim): Stock aliquots of 10 mg/mL at –20°C.
2. Lysozyme (Boehringer Mannheim): Stock aliquots of 10 mg/mL at –20°C.
3. Ribonuclease (RNase) A (Boehringer Mannheim).
4. 10X TE buffer: 100 m*M* Tris-HCL, pH 8.0, and 10 m*M* ethylenediaminetetraacetic acid (EDTA) in distilled water. For 1X TE, dilute in a ratio of 1:10 with distilled water.
5. 10% sodium dodecyl sulfate (SDS): Dissolve 10 g of SDS in 100 mL distilled water at 65°C.
6. 10% *N*-acetyl-N_3-trimethyl ammonium bromide (CTAB) (Merck) in 0.7*M* NaCl: Slowly add 10 g of CTAB to 4.1 g NaCl in 100 mL distilled water at 65°C. Adjust volume to 100 mL.
7. Chloroform/isoamyl alcohol (24:1): Mix 24 volumes of chloroform with 1 volume of isoamyl alcohol.
8. 5*M* NaCl: Dissolve 29.2 g NaCl (Fisher) in 100 mL distilled water.
9. 70% ethanol.
10. Water bath (80°C) and thermomixer (up to 60°C).
11. Centrifuge for 1.5-mL Eppendorf tubes.
12. Speed Vac (Savant Speed Vac Systems, GMI Inc., MN).

2.3. Electrophoresis

1. Electrophoretic-grade agarose (FMC).
2. 10X TBE buffer: Dissolve 108 g Tris-base and 55 g boric acid in 700 mL distilled water and 40 mL of 0.5*M* EDTA at pH 8.0. Adjust to 1 L and autoclave. For use, dilute to 1X in distilled water.
3. Horizontal electrophoretic chamber, 20 cm in length. Bio-Rad DNA Sub Cell chamber for membranes 20 × 15 cm (Bio-Rad Laboratories).
4. Electrical power supply (e.g., EC-103, E-C Apparatus Corp., Pittsburg, PA; EB103, Fisher Biotech, Pittsburgh, PA).

5. Loading dye: 5 mL 10X TBE, 25 mL glycerol, 15 mL H_2O, 5 mL 1% (w/v) double dye (1% bromophenol blue and xylene cyanol in H_2O).
6. DNA fragment size standards. Several size standards have been used. The most commonly used standard is Mt14323, which consists of chromosomal DNA isolated from clinical strain 14323, digested with *Pvu*II and loaded on each agarose gel (*see* **Note 1**).

2.4. Vacuum Blotting

1. Hybond-N+ (Nytron) membrane 20 × 15 cm (Amersham Biosciences, UK) or Zeta-Probe Blotting membranes (Bio-Rad Laboratories, USA).
2. 20X SSC stock solution: 3*M* NaCl, 0.3*M* Na_3-citrate, pH 7.0. Dilute 10X SSC in a ratio of 1:2 with distilled water.
3. 1*M* HCl (hydrogen chloride): Dilute 85.5 mL of concentrated HCl in 914.5 mL distilled water. For 0.25*M* HCl, dilute 1*M* HCl in a ratio of 1:4 with distilled water or 5 mL of concentrated HCl diluted to 500 mL in distilled water.
4. 4*M* NaOH (sodium hydroxide): Dissolve 160 g NaOH in 800 mL distilled water. Adjust to 1 L. For 0.4*M* NaOH, dilute 4*M* NaOH in a ratio of 1:10 with distilled water.
5. Soak I: 0.5*M* NaOH and 1.5*M* NaCl (dissolve 60 g NaOH and 262.98 g NaCl in 3 L of water).
6. Soak II: 0.5*M* Tris–HCl and 1.5*M* NaCl (dissolve 125.2 g Tris and 175.33 g NaCl in 2 L of water, adjust to pH 7.2 with 80 mL of HCl).
7. Vacuum blotter (VacuGene XL, Farmacia) (*see* **Note 2**).
8. UV crosslinker (FB-UV XL-100, Fisher Scientific, Fisher Biotech).

2.5. Preparation of Probe from Genomic DNA or Plasmid

1. The primers RIS3, 5′-CGTCGAACGGCTGATGACCA; 6110-R, 5′-GGCGGGTCCAGATGGCTTGC are used for the amplification of the IS*6110* probe from chromosomal *M. tuberculosis* DNA. T7 universal primers can be used for the amplification of the IS*6110* fragment from pUCIS.
2. A 50 µL reaction mixture should contain 10 pmol of each primer, 1 ng of genomic DNA, 200 µ*M* deoxyribonucleoside triphosphates, 1X PCR buffer (pH 8.3), 1.5 m*M* $MgCl_2$, and 1–2 units of *Taq* polymerase. The PCR conditions include 95°C at 4 min for initial denaturation followed by 35 cycles of denaturation at 95°C for 30 s, annealing at 62°C for 30 s, and extension at 72°C for 30 s followed by a single 2-min extension at 72°C.

2.6. Hybridization, Washing, and Detection

1. Enhanced Chemiluminescense Direct Nucleic Acid Labeling and Detection Kit (ECL kit, Amersham). The ECL kit includes labeling reagent, glutaraldehyde, double-distilled water for probe dilution, hybridization buffer, and blocking agent.

2. Pro-Blot hybridization oven (Labnet International Inc.) and roller bottles (Robbins Scientific or Labnet International Inc.).
3. Primary wash buffer: Mix 360 g urea and 4 g SDS (or 20 mL of 20% SDS) in 25 mL 20X SSC; adjust to 1 L with distilled water.
4. Secondary wash buffer: 2X SSC. Dilute 20X SSC stock solution at a ratio of 1:10 with distilled water.
5. Hyperfilm™ MP (Amersham Bioscience, UK) or Blue Lite Autorad Film 8 × 10 in. (ISC BioExpress, USA) or any appropriate sensitive films.
6. Film developing: Any appropriate equipment for X-ray film developing (alternatively, use manually mixed reagents for developing and fixation).

2.7. Data Analysis

1. Scanner.
2. Sun Sparc5 Workstation (Sun Microsystems) with Whole Band Analyzer software, version 3.4 (BioImage) *(13)*.
3. BioNumerics (Applied Math, Belgium, latest version 5.1), GelCompare (Applied Math, Belgium). Instructions are included in manual, or for more details *see* **ref. *14***.

3. Methods

3.1. Bacterial Cultures

1. Incubate LJ cultures for 5–8 wk at 37°C with caps slightly opened.
2. Collect bacteria before the color of LJ media turns yellow.
3. For 7H11 (7H10) agar cultures, place Petri dishes in ziplock bags to avoid drying (*see* **Note 3**).

3.2. Mycobacterium tuberculos*is DNA Isolation Procedure and Quantification*

1. Collect a loop-full of *M. tuberculosis* colonies from LJ slant or Middlebrook 7H11 agar plate (*see* **Note 4**) and suspend in 500 µL H_2O in a 1.5-mL Eppendorf tube.
2. For heat kill, incubate tubes immersed in a water bath for 30 min at 80°C.
3. Samples can be frozen at this stage at –20 or –70°C and stored, if necessary.
4. Place sample tubes on Eppendorf thermomixer adjusted to 60°C. Add 70 µL 10% SDS and 50 µL proteinase K (stock solution); mix for 1 h at 60°C in low mode with shaking.
5. Preheat 5*M* NaCl and 10% CTAB to 60°C.
6. While the samples are still at 60°C, add 100 µL 5*M* NaCl. Mix thoroughly by inverting by hand.

7. Add 100 µL 10% CTAB; mix thoroughly by inverting by hand.
8. Incubate further for 15 min at 60°C in the thermomixer. Freeze for 15 min at –70°C. Samples can be stored at that stage if necessary.
9. Defrost samples at 60°C in the thermomixer.
10. Add 700 µL chloroform/isoamyl alcohol (24:1); invert carefully by hand 20–25 times (do not shake). A white homogeneous solution should appear.
11. Centrifuge for 10 min at 16,000*g*.
12. Transfer the upper (aqueous) phase (~700 µL) to a new tube with 500 µL cold isopropanol; mix by tilting the tube up and down several times. DNA precipitate may be visible at this point.
13. Set at –20°C for at least 30 min or at 4°C overnight.
14. Centrifuge for 10 min at 16,000*g*.
15. Decant and wash the pellet with 70% ethanol. Centrifuge for 5–10 min at 16,000*g*.
16. Decant, dry in Speed Vac centrifuge for less than 10 min at low drying rate.
17. Add 55–100 µL H_2O (or 1X TE) depending on the size of DNA pellet. Run 5 µL on a 1% (w/v) agarose gel in TBE buffer for 1 h at approx 100 V with a 1-kb DNA ladder; stain gel in ethidium bromide solution for a few hours for better results.
18. The quantification of isolated chromosomal DNA can be achieved by visual evaluation of DNA concentration loaded on the gel (for example, *see* **Fig. 5**).

3.3. DNA Digestion and Electrophoresis

1. *Pvu*II restriction of genomic DNA: Add 2.5 µL buffer and 1.5 µL *Pvu*II to 21 µL of a total volume of DNA and water. Incubate in water bath for 4 h at 37°C.
2. Prepare 200–250 mL 1% agarose in 1X TBE buffer and pour gel (15 × 20 cm; 20-tooth comb for 18 samples and two flanking standards).
3. Following restriction, add 5 µL of loading dye to each sample, mix, and load gel. Load molecular weight markers or standards (STDs) in lanes 1 and 19 (if a second gel is run the same day, load STD in lanes 1 and 18 on the second gel to avoid confusion).
4. Run at 90 V until dye front has run approx 1 cm into the gel, then run overnight at 36 V at room temperature.
5. Stop gel when the dye front nears the end of the gel (approximately 16 h) and stain gel with ethidium bromide. Photograph the stained gel.

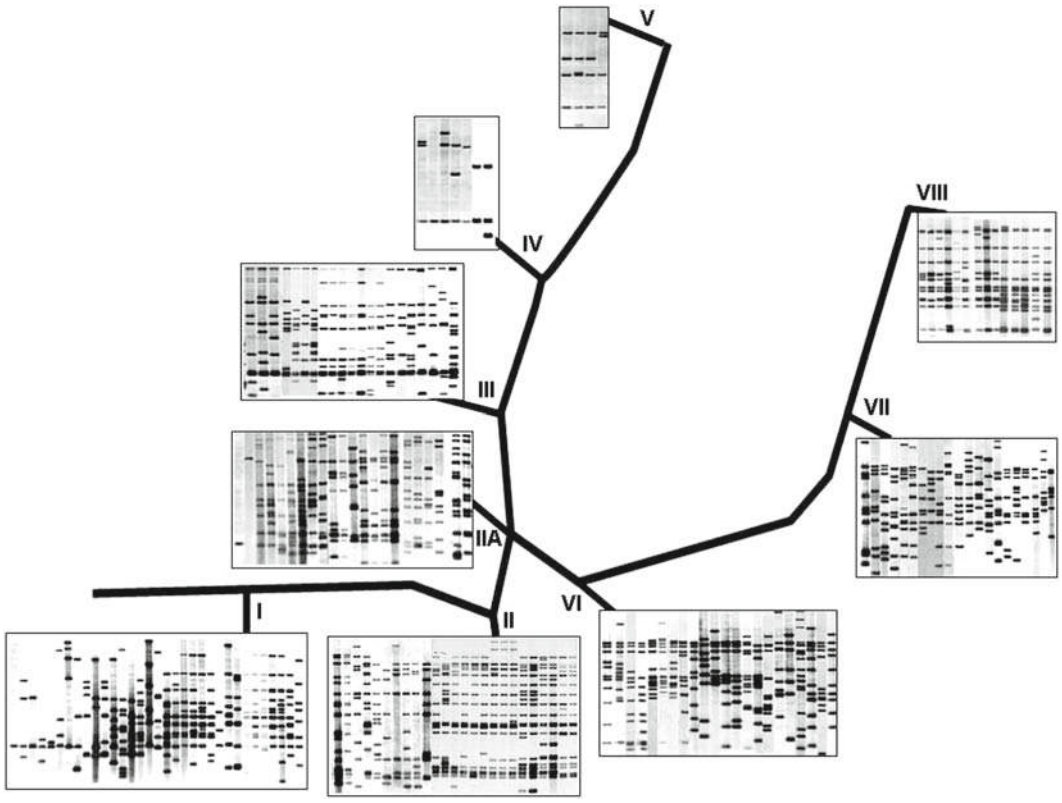


Fig. 5. Representative IS*6110* profiles superimposed on single-nucleotide polymorphism (SNP)-derived phylogenetic framework of *M. tuberculosis*. Based on SNP analysis of *M. tuberculosis* clinical isolates (including 1,743 strains from the Public Health Research Institute Tuberculosis Center strain collection), a phylogenetic tree with the nine clusters of *M. tuberculosis* isolates was used to illustrate common IS*6110* profiles. Some IS*6110* patterns are characteristic of given genetic cluster groups. (Adapted from **ref. *8***.)

3.4. Southern Blot

1. Rinse the Hybond-N+ membrane in water, followed by submersion in 10X SSC for 5 min.
2. Place the membrane on a porous support, cover with plastic mask, clamp unit, and place gel on top of the membrane. The size of the gel should be 0.5–1 cm larger than the window in the mask to create a vacuum in the transfer unit.
3. Attach to vacuum and adjust the suction unit to pull 50 mbar.
4. Flood the gel with 0.25*M* HCl for 20 min. Remove the fluid by vacuum aspiration using a pipet.
5. Flood the gel with soak I buffer for 20 min. Remove the fluid by vacuum aspiration using a pipet.
6. Flood the gel with soak II buffer for 20 min. Remove the fluid by vacuum aspiration using a pipet.
7. Flood the gel with 10X SSC for 1.5 h. Remove the fluid by vacuum aspiration using a pipet.

8. Without turning off the vacuum, remove the gel from the vacuum blotter, transfer the membrane to paper towels, and let the membrane air-dry for 10 min.
9. Irradiate the membrane twice to crosslink DNA fragments to membrane.

3.5. Hybridization

1. Prehybridize the membrane in 30 mL of hybridization buffer (ECL kit buffer with blocking agent, prepared in advance, as recommended by the manufacturer), rotating at approx 2.3 rcf and 42°C for 30 min in roller tube in the hybridization oven.
2. For each membrane, combine 10 µL of IS*6110*-specific probe and 10 µL of the provided water in an Eppendorf tube (*see* **Note 5**). Boil for 5 min in water and then place on ice for 10 min. Add an equal volume of labeling reagent (20 µL) and glutaraldehyde (ECL kit) (20 µL) (*see* **Note 6**), total volume of 60 µL per membrane. Incubate at 37°C in a water bath for 10–15 min.
3. Remove hybridization buffer from the tube, add the labeled probe into the buffer, and return to the tube containing the membrane.
4. Hybridize the membrane overnight, rotating at 3.3 rcf and 40°C.

3.6. Washing the Membrane After Hybridization

1. Remove the hybridization buffer, which can be reused once without adding additional probe if frozen at –20°C. Rinse the membrane in the roller bottle with approx 40 mL of primary wash buffer; discard the solution.
2. Add fresh primary wash buffer and set rotation of roller bottle for 30 min at approx 2.3 rcf at 40°C; discard the solution.
3. Rinse membrane with secondary wash buffer (2X SSC); discard solution. Add approx 30 mL secondary buffer and rotate the bottle for 10 min.
4. Remove the membrane from the bottle using 2X SSC buffer and soak in 2X SSC in a wide container on a shaking platform for 5–10 min.
5. Discard solution and repeat soaking as in **step 4** (the membrane should not be incubated in secondary wash buffer for more than 30 min).

3.7. Detection of Chemoluminescent-Labeled DNA Fragments

1. Mix 10 mL of solution 1 with 10 mL of solution 2 (ECL detection kit). This amount is sufficient for one membrane; when detecting more than one membrane, increase the total volume to 15 + 15 mL.
2. Incubate each membrane for 1 min at room temperature, carefully rotating the container. At this stage it is not required to work in darkness.
3. Wrap the membrane in plastic wrap, remove additional liquid with paper towels, and in a dark room place a film in the cassette (appropriate for X-ray-sensitive films).

4. Expose for approximately 10 min, replace film in dark room, and determine length of next exposure according to intensity of hybridization bands on the first film; increase time of exposure if bands are weak, and decrease exposure if bands are very dark. The time of exposure varies from 1 to 30 min, depending on the amount of chromosomal DNA loaded, quality of the hybridization probe, and ECL kit.

3.8. Rehybridization of the Membrane with Other Probes

The membrane used for IS*6110* RFLP can be reused for hybridization with other probes, such as the 5′-IS*6110* fragment or any other region of interest (*see* **refs. *9*** and ***12*** for examples). When the ECL kit is used, no additional stripping procedure is required. Simply expose the membrane to daylight for 20–24 h and prepare a new labeled probe as indicated. Prehybridization is not needed for the second probe. The same membrane can be rehybridized with different DNA probes up to six times without losing the quality of images.

3.9. Digitalization of Image

All software programs available for RFLP image analysis need the hybridization blot scanned and transformed into TIFF or JPEG file formats. The details of the RFLP analysis procedure are described by the software manufacturers and are not the subject of this chapter. In general, the image of the new isolate is compared to previously analyzed images in the collection, and the level of strain similarity (typically expressed as a percentage) is determined using statistical methods. Two images are considered identical if the number of hybridization bands and their location on the blots match 100%. Strains with (n + 1) hybridization bands may represent related strains, possibly from the same transmission chain, but further analysis using different biomarkers might be required for confirmation. Strains that differ by more than one hybridization band or strains showing a shift of some bands on the blot (different molecular weight of fragments) may still represent isolates from the same strain families but with a lower index of similarity. **Figure 4** represents selected images of three *M. tuberculosis* families: Haarlem, W-Beijing, and Latin America Mediterranean (LAM). Identical strains with IS*6110* copy number equal to or less than six require additional analyses (e.g., spoligotyping or MIRU-VNTR analysis).

4. Notes

1. The advantages of Mt14323 are easy amplification, wide size range of *Pvu*II-fragments (from 0.9 to 14 kb), and ability for direct interlaboratory strain comparisons (**Fig. 1**, lane 2).

The Centers for Disease Control and Prevention have developed a standard set of cloned IS*6110*-3′-fragments, ranging in size from 0.7 to 15 kb with 1-kb increments (**Fig. 1**, lane 1). This IS*6110* standard is used by all participating laboratories of the National Tuberculosis Genotyping and Surveillance Network, USA.

2. Alternative equipment for the capillary transfer can be used.
3. Fresh cultures of clinical *M. tuberculosis* isolates provide the best results for IS*6110*-based RFLP genotyping. LJ media are preferable for culture of *M. tuberculosis* strains as chromosomal DNA is less degraded, and hybridization bands are more distinguishable.
4. Use of prewetted cotton swabs makes this procedure more effective.
5. Depending on the probe concentration, 15 μL of the probe may be combined with 5 μL H_2O.
6. Thoroughly mix labeling reagent and DNA prior to addition of the glutaraldehyde.

References

1. World Health Organization (2007). Tuberculosis Facts. Available at: http://www.who.int/tb/publications/2007/factsheet_2007.pdf.
2. Zignol, M., Hosseini, M. S., Wright, A., Weezenbeek, C. L., Nunn, P., Watt, C. J., et al. (2006). Global incidence of multidrug-resistant tuberculosis. J. *Infect. Dis.* **194**, 479– 485.
3. Mahillon, J., and Chandler, M. (1998). Insertion sequences. Microbiol. *Mol. Biol. Rev.* **62**, 725–774.
4. McAdam, R. A., Hermans, P. W., van Soolingen, D., Zainuddin, Z. F., Catty, D., van Embden, J. D., et al. (1990). Characterization of a *Mycobacterium tuberculosis* insertion sequence belonging to the IS3 family. Mol. *Microbiol.* **4**, 1607–1613.
5. McEvoy, C. R., Falmer, A. A., Gey van Pittius, N. C., Victor, T. C., van Helden, P. D., and Warren, R. M. (2007). The role of IS*6110* in the evolution of *Mycobacterium tuberculosis. Tuberculosis* (Edinburgh, Scotland) **87**, 393–404.
6. Dale, J. W., Tang, T. H., Wall, S., Zainuddin, Z. F., and Plikaytis, B. (1997). Conservation of IS*6110* sequence in strains of *Mycobacterium tuberculosis* with single and multiple copies. Tuber. *Lung Dis.* **78**, 225–227.
7. van Embden, J. D., Cave, M. D., Crawford, J. T., Dale, J. W., Eisenach, K. D., Gicquel, B., et al. (1993). Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J. *Clin. Microbiol.* **31**, 406–409.
8. Mathema, B., Kurepina, N. E., Bifani, P. J., and Kreiswirth, B. N. (2006). Molecular epidemiology of tuberculosis: current insights. *Clin. Microbiol. Rev.* **19**, 658–685.
9. Kurepina, N., Likhoshvay, E., Shashkina, E., Mathema, B., Kremer, K., van Soolingen, D., et al. (2005). Targeted hybridization of IS*6110* fingerprints identifies the W-Beijing *Mycobacterium tuberculosis* strains among clinical isolates. *J. Clin. Microbiol.* **43**, 2148–2154.
10. Bifani, P. J., Plikaytis, B. B., Kapur, V., Stockbauer, K., Pan, X., Lutfey, M. L., et al. (1996). Origin and interstate spread of a New York City multidrug-resistant *Mycobacterium tuberculosis* clone family. *JAMA* **275**, 452–457.
11. Bifani, P. J., Mathema, B., Liu, Z., Moghazeh, S.L., Shopsin, B., Tempalski, B., et al. (1999). Identification of a W variant outbreak of *Mycobacterium tuberculosis* via population-based molecular epidemiology. *JAMA* **282**, 2321–2327.

12. Bifani, P., Mathema, B., Campo, M., Moghazeh, S., Nivin, B., Shashkina, E., et al. (2001). Molecular identification of streptomycin monoresistant *Mycobacterium tuberculosis* related to multidrug-resistant W strain. *Emerg. Infect. Dis.* **7**, 842–848.
13. Mathema, B., Bifani, P. J., Driscoll, J., Steinlein, L., Kurepina, N., Moghazeh, S. L., et al. (2002). Identification and evolution of an IS*6110* low-copy-number *Mycobacterium tuberculosis* cluster. J. *Infect. Dis.* **185**, 641–649.
14. Heersma, H. F., Kramer, K., and van Embden, J. (1998). Computer analysis of IS*6110* RFLP patterns of *Mycobacterium tuberculosis*, in *Mycobacteria Protocols*, Vol. 101, (Parish, T., and Stoker, N. G., eds.), Humana Press, London, pp. 395–422.

Chapter 15

spa Typing for Epidemiological Surveillance of *Staphylococcus aureus*

Marie Hallin, Alexander W. Friedrich, and Marc J. Struelens

Abstract

The *spa* typing method is based on sequencing of the polymorphic X region of the protein A gene (*spa*), present in all strains of *Staphylococcus aureus*. The X region is constituted of a variable number of 24-bp repeats flanked by well-conserved regions. This single-locus sequence-based typing method combines a number of technical advantages, such as rapidity, reproducibility, and portability. Moreover, due to its repeat structure, the *spa* locus simultaneously indexes micro- and macrovariations, enabling the use of *spa* typing in both local and global epidemiological studies. These studies are facilitated by the establishment of standardized *spa* type nomenclature and Internet shared databases.

Key words: Epidemiology; methicillin resistance; phylogeny; *S. aureus*; sequence analysis; software; staphylococcal protein A; typing methods.

1. Introduction

1.1. Typing Methods Available for Staphylococcus aureus

Staphylococcus aureus is a leading human pathogen responsible for a wide range of diseases, from superficial skin infections to life-threatening conditions, such as bacteremia, endocarditis, pneumonia, or toxic shock syndrome (*1*). Since the early 1960s, when they first emerged (*2*), strains of *S. aureus* resistant to methicillin and other β-lactams (MRSA) have spread worldwide and caused outbreaks in the hospital setting as well as in the community, thereby becoming a major public health threat (*3*). During the last decades, diverse typing methods, first phenotypic, then genotypic,

D.A. Caugant (ed.), *Molecular Epidemiology of Microorganisms*, Methods in Molecular Biology, Vol. 551
DOI: 10.1007/978-1-60327-999-4_15,

have been used for monitoring *S. aureus* spread. Among these, pulsed-field gel electrophoresis (PFGE) of genomic macrorestriction fragments is considered the gold-standard method (***4***). However, PFGE is a technically demanding and labor-intensive method. Moreover, its interpretation leaves room to subjectivity (***5***), and interlaboratory results comparison remains difficult and subject to strict adherence to standardized protocols and interpretation criteria (***6–8***).

Multilocus sequence typing (MLST), based on the sequence polymorphism of approx 500-bp long fragments of seven housekeeping genes was designed to study the *S. aureus* population genetic structure. This technique, applied to large *S. aureus* strain collections, revealed that the population structure is essentially clonal, and that the large majority of epidemic MRSA clones belong to a few phylogenetically distinct lineages or clonal complexes (CCs) (***9***).

MLST has also proved to be adequate for long-term global epidemiology and the study of recent evolution of *S. aureus* (***9,10***). However, MLST typing remains too expensive and labor intensive for its application to outbreak investigations and routine surveillance (***10,11***).

In recent years, more focused sequence-based methods have been developed to provide fast, unambiguous, and exportable typing data. Among these, the sequence determination of the polymorphic X region of *spa* gene, called *spa* typing, is gaining favor as a reliable tool for typing *S. aureus*. Frenay et al. were, in 1994, the first team to target the polymorphic X region of *spa* gene as an epidemiological marker. At that time, the X region was amplified and its size estimated by electrophoresis (***12***). In 1996, the same team improved the technique by performing sequence analysis of the X region (***13***). Since then, many studies have evaluated the usefulness of this technique for diverse epidemiological purposes and confirmed its ease of use and speed. Initially, two limitations hampered use of *spa* typing for surveillance: the lack of software capable of identifying and clustering repeat units and profiles and the lack of consensus nomenclature allowing interlaboratory exchange of results. These limits have been recently overcome, making *spa* typing a prime alternative to PFGE for typing *S. aureus*. In this chapter, we outline the biological basis of *spa* locus polymorphism and performance of *spa* sequence typing and describe the methods of analysis and data interpretation as well as international *spa* typing networks.

1.2. Structural Specificity of the spa *Gene Hypervariable X Region*

Protein A is a cell-wall component bound to the peptidoglycan of *S. aureus* by its COOH-terminal part. It interacts with the Fc-fragment of immunoglobulins by its NH_2-terminal part and thereby inhibits phagocytosis by polymorphonuclear leucocytes. The *spa* gene is composed of an N-terminal region encoding four

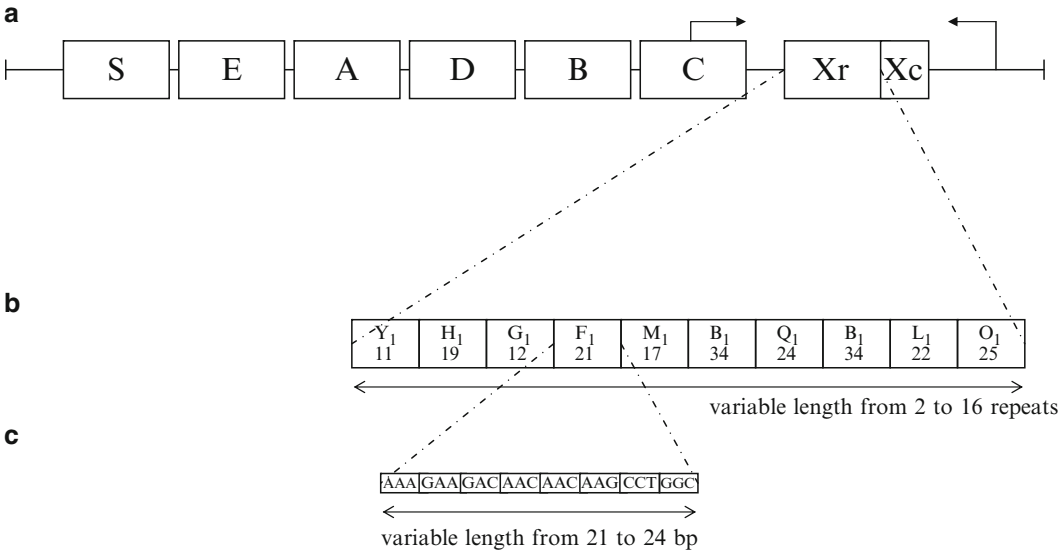


Fig. 1. (**a**) Schematic map of the *spa* gene. (Adapted from refs. *18, 20, 41*.) S is a signal sequence; A to D are IgG-binding domains; X is the C-terminal part, divided in two regions, the VNTR region (Xr) and a constant region coding for cell wall attachment (Xc). Arrows indicate the primers' localization. (**b**) The repeat structure of the Xr region. The *spa* type illustrated is t008 (Ridom-Harmsen et al. nomenclature) or YHGFMBQBLO (Kreiswirth nomenclature). (**c**) The DNA sequence of the *spa* repeat 21 (Ridom) or -F_1 (Kreiswrirth) repeat.

to five homologous immunoglobulin G (IgG) binding units, while its C-terminal sequence, called the X region, exhibits a variable number of short (24-bp) repeated units flanked by well-conserved regions (**Fig. 1**) (*14*).

The origin of the so-called variable number of tandem repeats (VNTR) structure of the X region is explained by the slipped strand mispairing model. Basically, illegitimate basepairing due to stretches and loops in short repeated unit motifs occurs during DNA replication, leading the DNA polymerase to delete or insert repeat units (*15*).

This VNTR structure is specific to neither the *spa* gene nor *S. aureus*. These hypervariable regions have been identified in many bacterial species and are frequently used as epidemiological markers. For example, the *spa* gene is one of the targets used by several multilocus VNTR analysis (MLVA) typing systems developed for *S. aureus*, together with *sdrC, D*, and *E, sspA, ClfA* and *B*, *cna*, or *fnBP* genes (*16*).

1.3. spa Typing Nomenclature

Two *spa* type designations have been used in the last years, one developed by Harmsen et al. (*17*) and one developed by Kreiswirth et al. (*18*). The latter has been changed recently, and comparison between the old designation and the new designation as well as a comparison between the Harmsen and the Kreiswirth nomenclature is only possible via computerized tools. The general

approach is, however, similar for both nomenclatures: Each repeat identified is associated to a code (numerical for the Harmsen et al. and alphanumerical for the Kreiswirth et al. nomenclature). Each *n*-long repeat profile corresponds to an *n*-long code constituted by the succession of the repeat's codes, as illustrated in **Fig. 1B**. In addition, in the Harmsen et al. nomenclature, a "type" number preceded by the letter t is then assigned to every distinct repeat profile (*see* **Note 1**). Due to its broader international use, we focus in the following section on the Harmsen et al. (Ridom) spa type designation.

1.4. spa *Typing Performance and Application Field*

As a sequence-based method, *spa* typing possesses many obvious advantages, such as rapidity, ease of use, suitability for computerized analysis, storage, and (ex)portability of results (***19***).

1.4.1. Typeability, Reproducibility, and Stability

Typeability is virtually 100%, although mutations in the flanking conserved regions of the X region, used for primer design, have occasionally been described (***20***), leading to amplification problems. However, a limitation of the clustering analysis may occasionally occur with isolates presenting short repeat profiles. Indeed, the epidemiological information contained by these profiles may be insufficient to permit reliable clustering (*see* **Note 2**).

Reproducibility is excellent (100%) both intra- and interlaboratory (***18,21***). The *spa* typing has been proved stable in vitro (***18***) and to a lesser extent in vivo (***13***). However, the longitudinal study of ten persistent infections of cystic fibrosis (CF) patients by a single *S. aureus* clone (as determined by PFGE) demonstrated occasional mutational events (deletions, point mutation, or duplication of *spa* repeats) in 10% of the sequential isolates studied (***22***). This finding could also be explained by the existence of closely related co-colonizing isolates, as described later by the same team (***23***), and cannot be generalized given the peculiarity of the ecological niche and selective pressure present in the airways of CF patients.

1.4.2. Epidemiological Concordance and Discriminatory Power

The use of *spa* typing for outbreak investigation was validated by Shopsin et al. (***18***) on a well-documented strain collection containing 29 isolates belonging to four distinct outbreaks. Discriminatory power of *spa* typing, evaluated on several large strain collections, was found to be similar to PFGE, with a Simpson Index of diversity ranging from 0.97 to 0.98 (***4,20,24***).

1.4.3. Concordance with Gold Standard Methods and Phylogenetic Inference

Besides a high discriminatory power, *spa* typing is usually in good concordance with PFGE (using the Tenover criteria (***25***)), either at the type level (from 96% to 98%) (***20,24***) or between clusters (93%, using BURP (based upon repeat patterns) algorithm; *see* **Subheading 3.4.2.)** (***4,24,26***). Furthermore, concordance with MLST and BURST (based upon related sequences) clustering also proved to be very high (97% to 99%) (***4,24***).

This ability of *spa* typing to combine a high discriminatory power with a high concordance with MLST as a single-locus marker resides in its repeat composition and organization. Point mutations (occurring at a low rate and subject to purifying selection) permit a reliable lineage assignment, while additions or deletions of repeats (fast-occurring) index intralineage variations, enabling the use of *spa* typing for both long- and short-term epidemiology (***19***).

However, when assuming that an isolate belongs to an MLST lineage based on its *spa* type, one should be aware of possible misclassification problems, as identified for a few lineages by several authors (***4,24,27,28***). Indeed, strains belonging to distant MLST CCs can present identical or similar *spa* profiles and cluster (using the BURP algorithm) together in a unique *spa* group. This still unexplained phenomenon could possibly be in certain cases due to recombination events involving the *spa* locus (*27*). Other cases can be caused by large chromosomal replacement encompassing the *spa* gene, such as described by Robinson and Enright for ST239 and ST34 (*28*).

2. Materials

The materials listed are only suggestions as many excellent alternatives exist.

2.1. DNA Extraction

1. Columbia agar plates with 5% sheep blood (Biomérieux, Marcy l'Etoile, France).
2. Lysostaphin solution (1 mg/mL) (AMBI Products LLC, New York).
3. Proteinase K solution (1 mg/mL) (Sigma).
4. DNase (deoxyribonuclease)-free water.
5. Tris-HCl, 0.1*M*, pH 8.0 solution.
6. Water baths or heater blocks (37, 60, 100°C).
7. Vortex.

2.2. Polymerase Chain Reaction (PCR)

1. AmpliTaq DNA polymerase (Applied Biosystems, Foster City, CA).
2. Deoxynucleotide 5′-triphosphates (dNTPs; Promega, Madison, WI).
3. High-performance liquid chromatographic-cleaned primers (spa-1113f, spa-1514r; for complete sequence, *see* **Subheading 3.2.**) (MWG-Biotech, Ebersberg, Germany).
4. PCR buffer II (Applied Biosystems).

5. End-point thermocycler: GeneAmp PCR System 9700 (Applied Biosystems).
6. Elution spin column: Kit Quantum prep PCR Kleen Spin Columns (Bio-Rad).

2.3. DNA Sequencing

1. ABI Prism BigDye Terminator V3.1 Sequencing Kit (Applied Biosystems).
2. ABI prism 4100 sequencing machine (Applied Biosystems).
3. Microplates: MultiScreen HV, clear plate 45UL (Millipore, Billerica, MA).
4. Sephadex G 50 (Amersham Biosciences, Freiburg, Germany).
5. Microplate centrifuge, up to 2100 g, rotor 11123 (Sigma).

2.4. Data Analysis

1. StaphType software (Ridom GmbH, Würzburg, Germany) or Bionumerics (Applied Maths, Ghent, Belgium).
2. Internet connection.

3. Methods

3.1. DNA Extraction

Multiple extraction protocols ranging from a simple boiling step (*29*) to commercial tissue or blood extraction kits (*27*) or glass bead mechanical lysis have been described to be suitable for *spa* typing (*see* **Note 3**). A rapid extraction protocol can be used as follows:

1. Suspend one colony of *S. aureus* cultured for 24 h on Columbia agar with 5% sheep blood in 45 µL of DNase-free water and 5 µL of lysostaphin (1 mg/mL), vortex, and incubate for 10 min at 37°C.
2. Add 45µL of DNase-free water, 5 µL of proteinase K 2 mg/mL and 150 µL of Tris-HCl 0.1*M,* pH 8.0; vortex and incubate 10 min at 60°C and then 5 min at 100°C.
3. Of this lysate, 5 µL is used as the DNA template in the PCR reaction.

3.2. DNA Amplification

Several pairs of primers have been described, numbered from the forward strand of *S. aureus* DNA (GenBank accession no. J01786), for example:

spa-1113f [1092–1113] (5′-TAA AGA CGA TCC TTC GGT GAG C) and spa-1514r [1534–1514] (5′-CAG CAG TAG TGC CGT TTG CT-3′) (***30***).

spa-1095f [1095–1113] (5′- AGA CGA TCC TTC GGT GAG C) and spa-1517r [1517–1496] (5′-GCT TTT GCA ATG TCA TTT ACT G-3′) (***17,18***).

The DNA amplification protocol (recommended by the Ridom software supplier, www.ridom.de) is the following:

1. Add genomic staphylococcal DNA in a PCR mixture to achieve 50 µL of final volume containing 1.25 units of *Taq* polymerase, 1.5 m*M* $MgCl_2$, 200 µ*M* dNTPs, 0.2 µ*M* of each primer (spa-1113f, spa-1514r) and 5 µL of 10X PCR buffer.
2. Cycling conditions consist of an initial denaturation step of 5 min at 80°C, followed by 35 cycles of 45 s of denaturation at 94°C, 45 s of annealing at 60°C, 90 s of extension at 72°C, and a final extension step of 10 min at 72°C (*see* **Notes 4–6**).
3. The PCR product is then purified, either by an enzymatic method or by elution spin column, and can be stored at 4°C.

3.3. DNA Sequencing

1. Use the following reaction mix: 20 to 30 ng of amplified and purified DNA in a final reaction volume of 10 µL containing 2 µL of premix and 1 µL of buffer from the kit and 0.5 m*M* of each primer (use the same primers as for the amplification step).
2. Amplification parameters are the following: An initial denaturation step of 2 min at 96°C, followed by 25 cycles of 30 s of denaturation at 96°C, 15 s of annealing at 50°C, and 60 s of extension at 60°C.
3. The products are then purified and concentrated prior to sequencing, either by ethanol precipitation or by commercial elution spin columns (Dye-ex, Quiagen, Hilden, Germany) or microplates loaded with Sephadex G 50.

3.4. Data Interpretation

3.4.1. spa *Type Assignment*

The VNTR structure of the *spa* locus makes the traditional sequence alignment (using the substitutions, insertions, and deletions [indel] of a single-position model) improper to accurately identify *spa* repeat units and assess how these units are organized. Several programs, both "in-house" and commercial, have been described and can be successfully used for this purpose (*17,18,31*). Among those, the StaphType software (Ridom) is the most widely used. Other bioinformatic tools, such as Bionumerics (Applied Maths) also allow *spa* analysis and *spa* type designation using the Ridom nomenclature. Submissions from the Bionumerics software to the *spa* server will be possible in the near future via the online SeqNet gatekeeper interface (www.seqnet.org; *see* **Subheading 3.4.3.**). Both software include internal quality control systems that ensure that extrapolation of repeat and repeat organization are made on sequence data (chromatograms) of sufficient quality (*see* **Note 7**).

3.4.2. *Grouping of Related* spa *Types*

Until recently, the only way to cluster or classify *spa* types was to visually estimate the similarity of their repeat profiles. This method, although feasible with limited size strain collections and found to concordantly classify strains as compared with MLST

results (***20,29,32,33***), is difficult to apply to large collections. The BURP algorithm is an automated algorithm—implemented in the StaphType software—that can cluster (*spa*-CC) *spa* types (***17,34***). Repeat duplication and excision are taken into account (in addition to single-position substitution and indel events) when the relatedness of different *spa* types is calculated. A "cost" accounting for the "steps" of evolution between each examined pair of *spa* types is calculated, whereas the algorithm tries to minimize these steps (parsimony assumption) (***34***). BURP offers two user-defined parameters that influence clustering: exclusion of *spa* types that are shorter than *x* repeats and the maximum number of costs *y* for clustering *spa* types into the same group. Using these parameters, short *spa* types (presenting limited evolutionary information) can be excluded from further analysis (*see* **Note 2**), and maximum costs can be adapted to the size of the strain collection studied and the question asked in terms of space and time evolutionary scale (small versus large outbreak investigation versus long-term epidemiological surveillance studies) (*see* **Note 8**).

3.4.3. The StaphType Software

The StaphType software combines three modules: a sequence editor, a database, and a report generator. For each *spa* sequence downloaded by the software, epidemiological information concerning the isolate typed can be recorded in the database module.

3.4.3.1. Sequence Analysis and spa *Type Designation*

The sequence analysis starts with the download of both forward and reverse sequences files (FASTA format or preferably ABI and SCF chromatograms). The software searches then automatically for the 5′ and 3′ signature sequences (conserved flanking regions), constructs a consensus sequence, and detects the *spa* repeats succession (***17***). In case of already known *spa* repeat succession, the *spa* type designation is automatically downloaded from the *spa* server Web site. New *spa* repeats and *spa* types detected by one laboratory using the StaphType software are automatically given a preliminary name in the local database (e.g., txAA or txAB). The laboratory typing data can then be synchronized via the Internet; the new sequences are then matched automatically with *spa* types found by other participants. If the repeat succession is revealed as new, a new type number is assigned for all future detection of this *spa* repeat profile, ensuring a continuously updated common nomenclature. Type numbers are assigned by the order of submission; no relatedness can be deduced from the closeness of two t-numbers. If the repeat succession has already been described and synchronized by another laboratory, the preliminary name in the local database of the inquiring laboratory is automatically changed to the preexisting denomination.

3.4.3.2. *Automated Quality Control of* spa *Typing Data*

As the *spa* server receives more than 1,000 submissions per month, a major goal is the maintenance of excellence of the data

quality. Therefore, the curator of the SeqNet.org database (*see* **Subheading 3.4.5.**) has set up rules for procedure and internal and external quality control schemes:

1. An internal quality control system is integrated in the StaphType software: To each downloaded sequence is attached a quality index, which corresponds to a sequence error probability. The *spa* typing sequences with low reliability cannot be synchronized via the server and are rejected by the SeqNet curator (*see* **Note 9**).
2. The external quality control consists of the performance of a certification for all new SeqNet.org members and a regular proficiency test, based on known ring trials. This external quality certificate is sent out to the laboratory when capacities for high-quality sequence typing has been established. During the certification process, the curators assist the new SeqNet.org aspirants in the development of sequence capacity, often accompanied by a 3-d stay at the sequencing facilities of the coordinators or participating in one of the hands-on laboratory workshops.

3.4.3.3. Data Ownership

Data ownership on the *spa* server is ruled by the SeqNet.org initiative, which curates the data for all submitters. It is important to mention that all data on the *spa* server are strictly incrementally synchronized. This means that all synchronized data, after passing quality control and assignment of the *spa* type, are stored with a single laboratory identifier. Every submitter using direct submission can choose which epidemiological data should be shared on the Web site (*see* **Note 10**). International study groups or regional and national networks can opt for not making visible their data submission on the public home page as long as wished by the interested group. In this way, intellectual data property of each single submitter is protected.

3.4.4. The Central spa Server

The Harmsen et al. nomenclature for the designation of *spa* types has been made universally accessible by establishing the central *spa* server. It allows the automated quality control of submitted sequence data, and the central synchronization rends the submitted data publicly available on the online Web site (www.SpaServer.ridom.de). Users of other *spa* analyzing software tools than StaphType are able to synchronize with the *spa* server via an online uploading interface while fulfilling all given quality criteria checked by the SeqNet.org curators. Until now, agreements between SeqNet.org and two developers of *spa* analyzing software (Ridom at www.ridom.de and Applied Maths at www.applied-maths.com) have been achieved. SeqNet.org will serve as gatekeeper for quality for the synchronization of *spa* sequences from submitters using one of the *spa* analyzing tools.

3.4.5. The SeqNet.org spa Typing Network

The central *spa* server (which has been developed by Ridom) is curated by the SeqNet.org initiative (***35***) on behalf of all users. SeqNet.org currently is an initiative of 45 laboratories from 25 European countries (1 laboratory from Lebanon) founded in 2004 at the University of Münster in Germany (http://www.SeqNet.org). Its main objective is to establish a European network of excellence for sequence-based typing of microbial pathogens, having its main focus on *S. aureus*. SeqNet.org comprises a large number of national reference laboratories as well as university and some veterinary laboratories. The principle goal of SeqNet.org is to create unambiguous, electronic, portable, easily comparable typing data of excellent quality for local infection control and national and European surveillance of sentinel microorganisms, such as MRSA.

Currently, parallel to the SeqNet.org laboratories, more than 140 other submitting laboratories have synchronized their *spa* types with the database. Although, the *spa* database in its current form essentially is used as a *spa* type dictionary, ensuring a common nomenclature, providing molecular typing data in real time, and maintaining typing data quality, its data entries on frequencies of *spa* types and country of submission can already provide valuable information regarding geographical dissemination and occurrence of the *spa* types by country (***36***).

Furthermore, the *spa* server can be used by regional, national, or international public health or research networks to filter *spa* data from the network's participating laboratories, hospitals, and medical practices. Important examples are the Dutch-German cross-border networks EUREGIO MRSA-net Twente/Münsterland (***37,38***) and the MRSA network of the EUREGIO Maas-Rhein (***39***) In both cases, *spa* typing ensures not only the intrahospital but also the cross-border comparability and Euregional data ownership of the typing data.

4. Notes

1. A translating tool from one nomenclature to the other can be downloaded from the *spa* server Web site (www. SpaServer.ridom.de).
2. Isolates presenting short *spa* profile, although technically typable, should be excluded from clustering analysis. Five has been proposed (***34***) as the minimum number of repeats necessary to infer relatedness.
3. An initial "staphylococcal-specific" lysis step using lysostaphin (24,27) is however recommended to ensure sufficient bacterial lysis.

4. Several other amplification protocols have been described and proved to be efficient (***18,40***).
5. Visualization of the amplified DNA by conventional electrophoresis in 1 or 2% agarose gel is recommended prior to the sequencing step. The average amplified product size should be between 300 and 600 bp but varies following the number of *spa* repeats.
6. For isolates that are nontypable using the primers cited, SeqNet.org recommends using the following primers (A. Mellmann, personal communication 2007).
   (a) spa-239f (5′-ACTAGGTGTAGGTATTGCATCTGT-3′)
   (b) spa-1717r (5′-TCCAGCTAATAACGCTGCACCTAA-3′)
   (c) spa-1084f (5′-ACAACGTAACGGCTTCATCC-3′)
   (d) spa-1618r (5′-TTAGCATCTGCATGGTTTGC-3′)

   However, 1 of 1,000 isolates remains nontypable. The reason might be that some *S. aureus* isolates present large deletions in the *spa* gene that can affect the primer binding sites.
7. Manual editing of *spa* sequence data should be avoided as much as possible because it can easily lead to misidentification of repeats and subsequently to the attribution of an incorrect *spa* type.
8. This parameter is by default set to four by the Ridom software. However, this was calibrated to suit long-term evolution characterization (i.e., maximal concordance with MLST data) (***34***). For small outbreak investigation, this parameter can be lowered.
9. If the 5′/3′ signatures are found, the repeats are correct, and the sequence is traced correctly, the reliability value given is 100 (good). If 5′/3′ signatures are found, the repeat succession contains no low-quality basis, and there is a consensus of traces, then the reliability value for quality given is 110 (very good) (***12***). Last, if the criteria for 110 are fulfilled and there are fewer than five editing steps of the sequence, the reliability value for quality given is 120 (excellent). In the case that the 5′/3′ signatures are not correctly found, signature positions are shifted, or base quality is low, the reliability values are between 90 and 40 (sufficient) or between 30 and 0 (poor) (***33***). Each *spa* typing sequence with a reliability value lower than 100 cannot be synchronized via the server and is rejected by the SeqNet curators.
10. The submitter is also able to withdraw his or her data at any time by resynchronizing with the server and indicating the deletion of its submission. In such case, only the *spa* type and the information on the sequence quality will remain on the server.

References

1. Lowy, F. D. (1998). *Staphylococcus aureus* infections. *N. Engl. J. Med.* **339**, 520–532.
2. Jevons, M. P. (1961). "Celbenin"-resistant Staphylococci. *BMJ* **1**, 124–125.
3. Aires, d. S., and de Lencastre, H. (2004). Bridges from hospitals to the laboratory: genetic portraits of methicillin-resistant *Staphylococcus aureus* clones. *FEMS Immunol. Med. Microbiol.* **40**, 101–111.
4. Strommenger, B., Kettlitz, C., Weniger, T., Harmsen, D., Friedrich, A. W., and Witte, W. (2006). Assignment of *Staphylococcus* isolates to groups by *spa* typing, SmaI macrorestriction analysis, and multilocus sequence typing. *J. Clin. Microbiol.* **44**, 2533–2540.
5. Tenover, F. C., Arbeit, R., Archer, G., Biddle, J., Byrne, S., Goering, R., et al. (1994). Comparison of traditional and molecular methods of typing isolates of *Staphylococcus aureus*. *J. Clin. Microbiol.* **32**, 407–415.
6. Murchan, S., Kaufmann, M. E., Deplano, A., De Ryck, R., Struelens, M., Zinn, C. E., et al. (2003). Harmonization of pulsed-field gel electrophoresis protocols for epidemiological typing of strains of methicillin-resistant *Staphylococcus aureus:* a single approach developed by consensus in 10 European laboratories and its application for tracing the spread of related strains. *J. Clin. Microbiol.* **41**, 1574–1585.
7. Deplano, A., Schuermans, A., Van Eldere, J., Witte, W., Meugnier, H., Etienne, J., et al. (2000). Multicenter evaluation of epidemiological typing of methicillin-resistant *Staphylococcus aureus* strains by repetitive-element PCR analysis. The European Study Group on Epidemiological Markers of the ESCMID. *J. Clin. Microbiol.* **38**, 3527–3533.
8. Cookson, B. D., Aparicio, P., Deplano, A., Struelens, M., Goering, R., and Marples, R. (1996). Inter-centre comparison of pulsed-field gel electrophoresis for the typing of methicillin-resistant *Staphylococcus aureus*. *J. Med. Microbiol.* **44**, 179–184.
9. Enright, M. C., Day, N. P., Davies, C. E., Peacock, S. J., and Spratt, B. G. (2000). Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. *J. Clin. Microbiol.* **38**, 1008–1015.
10. Enright, M. C., Robinson, D. A., Randle, G., Feil, E. J., Grundmann, H., and Spratt, B. G. (2002). The evolutionary history of methicillin-resistant *Staphylococcus aureus* (MRSA). *Proc. Natl. Acad. Sci. U S A* **99**, 7687–7692.
11. Peacock, S. J., de Silva, G. D., Justice, A., Cowland, A., Moore, C. E., Winearls, C. G., et al. (2002). Comparison of multilocus sequence typing and pulsed-field gel electrophoresis as tools for typing *Staphylococcus aureus* isolates in a microepidemiological setting. *J. Clin. Microbiol.* **40**, 3764–3770.
12. Frenay, H. M., Theelen, J. P., Schouls, L. M., Vandenbroucke-Grauls, C. M., Verhoef, J., van Leeuwen, W. J., et al. (1994). Discrimination of epidemic and nonepidemic methicillin-resistant *Staphylococcus aureus* strains on the basis of protein A gene polymorphism. *J. Clin. Microbiol.* **32**, 846–847.
13. Frenay, H. M., Bunschoten, A. E., Schouls, L. M., van Leeuwen, W. J., Vandenbroucke-Grauls, C. M., Verhoef, J., et al. (1996). Molecular typing of methicillin-resistant *Staphylococcus aureus* on the basis of protein A gene polymorphism. *Eur. J. Clin. Microbiol. Infect. Dis.* **15**, 60–64.
14. Guss, B., Uhlen, M., Nilsson, B., Lindberg, M., Sjoquist, J., and Sjodahl, J. (1984). Region X, the cell-wall-attachment part of staphylococcal protein A. *Eur. J. Biochem.* **138**, 413–420.
15. van Belkum, A. (1999). The role of short sequence repeats in epidemiologic typing. *Curr. Opin. Microbiol.* **2**, 306–311.
16. van Belkum, A. (2007). Tracing isolates of bacterial species by multilocus variable number of tandem repeat analysis (MLVA). *FEMS Immunol. Med. Microbiol.* **49**, 22–27.
17. Harmsen, D., Claus, H., Witte, W., Rothganger, J., Claus, H., Turnwald, D., et al. (2003). Typing of methicillin-resistant *Staphylococcus aureus* in a university hospital setting by using novel software for *spa* repeat determination and database management. *J. Clin. Microbiol.* **41**, 5442–5448.
18. Shopsin, B., Gomez, M., Montgomery, S. O., Smith, D. H., Waddington, M., Dodge, D. E., et al. (1999). Evaluation of protein A gene polymorphic region DNA sequencing for typing of *Staphylococcus aureus* strains. *J. Clin. Microbiol.* **37**, 3556–3563.
19. van Belkum, A., Tassios, P. T., Dijkshoorn, L., Haeggman, S., Cookson, B., Fry, N. K., et al. (2007). Guidelines for the validation and application of typing methods for use in bacterial epidemiology. *Clin. Microbiol. Infect.* **13** (Suppl. 3), 1–46.
20. Koreen, L., Ramaswamy, S. V., Graviss, E. A., Naidich, S., Musser, J. M., and Kreiswirth, B. N. (2004). *spa* typing method for discriminating among *Staphylococcus aureus* isolates: implications for use of a single marker to detect genetic micro- and macrovariation. *J. Clin. Microbiol.* **42**, 792–799.

21. Aires-de-Sousa, M., Boye, K., de Lencastre, H., Deplano, A., Enright, M. C., Etienne, J., et al. (2006). High interlaboratory reproducibility of DNA sequence-based typing of bacteria in a multicenter study. *J. Clin. Microbiol.* **44**, 619–621.
22. Kahl, B. C., Mellmann, A., Deiwick, S., Peters, G., and Harmsen, D. (2005). Variation of the polymorphic region X of the protein A gene during persistent airway infection of cystic fibrosis patients reflects two independent mechanisms of genetic change in *Staphylococcus aureus*. *J. Clin. Microbiol.* **43**, 502–505.
23. Goerke, C., Gressinger, M., Endler, K., Breitkopf, C., Wardecki, K., Stern, M., et al. (2007). High phenotypic diversity in infecting but not in colonizing *Staphylococcus aureus* populations. *Environ. Microbiol.* **9**, 3134–3142.
24. Hallin, M., Deplano, A., Denis, O., de Mendonca, R., De Ryck, R., and Struelens, M. J. (2007). Validation of pulsed-field gel electrophoresis and *spa* typing for long-term, nationwide epidemiological surveillance studies of *Staphylococcus aureus* infections. *J. Clin. Microbiol.* **45**, 127–133.
25. Tenover, F. C., Arbeit, R. D., Goering, R. V., Mickelsen, P. A., Murray, B. E., Persing, D. H., et al. (1995). Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. *J. Clin. Microbiol.* **33**, 2233–2239.
26. Faria, N. A., Carrico, J. A., Oliveira, D. C., Ramirez, M., and de Lencastre, H. (2008). Analysis of typing methods for epidemiological surveillance of both methicillin-resistant and methicillin-susceptible *Staphylococcus aureus* strains. *J. Clin. Microbiol.* **46**, 136–144.
27. Strommenger, B., Braulke, C., Heuck, D., Schmidt, C., Pasemann, B., Nubel, U., et al. (2008). *spa* typing of *Staphylococcus aureus* as a frontline tool in epidemiological typing *J. Clin. Microbiol.* **46**, 574–581.
28. Robinson, D. A., and Enright, M. C. (2004). Evolution of *Staphylococcus aureus* by large chromosomal replacements*J. Bacteriol.* **186**, 1060–1064.
29. Ruppitsch, W., Indra, A., Stoger, A., Mayer, B., Stadlbauer, S., Wewalka, G., et al. (2006). Classifying *spa* types in complexes improves interpretation of typing results for methicillin-resistant *Staphylococcus aureus*. *J. Clin. Microbiol.* **44**, 2442–2448.
30. Mellmann, A., Friedrich, A. W., Rosenkotter, N., Rothganger, J., Karch, H., Reintjes, R., et al. (2006). Automated DNA sequence-based early warning system for the detection of methicillin-resistant *Staphylococcus aureus* outbreaks. *PLoS Med.* **3**, e33.
31. Oliveira, D. C., Crisostomo, I., Santos-Sanches, I., Major, P., Alves, C. R., Aires-de-Sousa, M., et al. (2001). Comparison of DNA sequencing of the protein A gene polymorphic region with other molecular typing techniques for typing two epidemiologically diverse collections of methicillin-resistant *Staphylococcus aureus*. *J. Clin. Microbiol.* **39**, 574–580.
32. Feil, E. J., Cooper, J. E., Grundmann, H., Robinson, D. A., Enright, M. C., Berendt, T., et al. (2003). How clonal is *Staphylococcus aureus*? *J. Bacteriol.* **185**, 3307–3316.
33. Robinson, D. A., and Enright, M. C. (2003). Evolutionary models of the emergence of methicillin-resistant *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* **47**, 3926–3934.
34. Mellmann, A., Weniger, T., Berssenbrugge, C., Rothganger, J., Sammeth, M., Stoye, J., et al. (2007). Based Upon Repeat Pattern (BURP): an algorithm to characterize the long-term evolution of *Staphylococcus aureus* populations based on *spa* polymorphisms. *BMC Microbiol.* **7**, 98.
35. Friedrich, A. W., Witte, W., Harmsen, D., de Lencastre, H., Hryniewicz, W., Scheres, J., et al. (2006). SeqNet.org: a European laboratory network for sequence-based typing of microbial pathogens. *Euro Surveill.* **11**, E060112.
36. von Eiff, C., Maas, D., Sander, G., Friedrich, A. W., Peters, G., and Becker, K. (2008). Microbiological evaluation of a new growth-based approach for rapid detection of methicillin-resistant *Staphylococcus aureus*. *J. Antimicrob. Chemother.* **61**, 1277–1280.
37. Friedrich, A., Daniels-Haardt, I., Gemert-Pijnen, J., Hendrix, M., von Eiff, C., Becker, K., et al. (2007). A regional network for the prevention and control of infections due to MRSA: EUREGIO MRSA-net Twente/Münsterland. *Epidemiol. Bull.* **33**, 307–311.
38. Daniels-Haardt, I., Verhoeven, F., Mellmann, A., Hendrix, M. G., Gemert-Pijnen, J. E., and Friedrich, A. W. (2006). [EUREGIO-projekt MRSA-net Twente/Munsterland. Creation of a regional network to combat MRSA]. *Gesundheitswesen* **68**, 674–678.
39. Deurenberg, R. H., Vink, C., Oudhuis, G. J., Mooij, J. E., Driessen, C., Coppens, G., et al. (2005). Different clonal complexes of methicillin-resistant *Staphylococcus aureus* are disseminated in the Euregio Meuse-Rhine region. *Antimicrob. Agents Chemother.* **49**, 4263–4271.
40. Oliveira, D. C., Crisostomo, I., Santos-Sanches, I., Major, P., Alves, C. R., Aires-de-Sousa, M., et al. (2001). Comparison of DNA

sequencing of the protein A gene polymorphic region with other molecular typing techniques for typing two epidemiologically diverse collections of methicillin-resistant *Staphylococcus aureus*. *J. Clin. Microbiol.* **39**, 574–580.

41. Uhlen, M., Guss, B., Nilsson, B., Gatenbeck, S., Philipson, L., and Lindberg, M. (1984). Complete sequence of the staphylococcal gene encoding protein A. A gene evolved through multiple duplications. *J. Biol. Chem.* **259**, 1695–1702.

Chapter 16

Sequencing of Viral Genes

Carol Holm-Hansen and Kirsti Vainio

Abstract

The use of molecular techniques in epidemiology gives a better understanding of viral transmission and diversity, and helps to define and characterizc outbreaks. By elucidating transmission patterns and defining outbreak parameters, appropriate preventive measures can be implemented in a timely fashion. Previously, the understanding of viral classification and phylogeny was difficult due to the challenges inherent in studying viruses. Automated cycle sequencing that uses fluorescently labeled nucleotides has revolutionized epidemiological studies of viruses at the molecular level. Sequencing of viral genes enables the identification and characterization of viruses, and sequence data are essential when investigating the etiology, dissemination, and transmission of viral infections as well as for disease surveillance and prevention. The present chapter focuses on the use of sequence analyses in epidemiological investigations.

Key words: Automated cycle sequencing, BigDye Terminator chemistry, PCR amplicon, phylogenetic analysis, virus diversity.

1. Introduction

The field of molecular epidemiology has emerged from the integration of molecular data into traditional epidemiologic research. Molecular techniques enable the characterization and comparison of different virus strains at the genomic level and are important tools for investigating the epidemiology of viral infections, at an individual or global level, and in retrospective investigations or surveillance. Sequence data have made it possible to distinguish and characterize local outbreaks, to detect dispersed international outbreaks, and to identify transmission chains by tracing the source of environmentally transmitted viruses (*1*) or

D.A. Caugant (ed.), *Molecular Epidemiology of Microorganisms*, Methods in Molecular Biology, Vol. 551
DOI: 10.1007/978-1-60327-999-4_16,

the dissemination routes of person-to-person (*2*) and zoonotic viral infections (*3*). Sequence data have also enabled the study of the origin and relatedness of viral strains, the detection of new strains and variants, and the emergence of drug resistance across strain generations. In addition, monitoring the diversity of viral agents in clinical virology contributes to better diagnostics, treatment, and prophylaxis (*4*). Combined molecular and epidemiological data have been used to improve prevention and control of infectious viral diseases and thus provide clear benefits for public health. Some examples for which sequence data have been used to complement epidemiological investigations are given below.

The application of sequence data has been essential for the surveillance of measles in the World Health Organization (WHO) European region. While the WHO European Region has targeted the elimination of measles by 2010, this goal may not be fully realized. Several measles genotypes imported from other continents have caused prolonged circulation and large outbreaks among unvaccinated and highly mobile communities in several European countries. In Norway, with only a few cases of measles per year, molecular typing has shown that all cases have been associated with import from endemic regions or specific reservoirs (*5–7*). As shown in **Fig. 1**, all sequences obtained from measles cases in Norway were identical to virus types circulating in countries where the patients came from or had recently traveled.

Combined sequence and epidemiological data have also given information on transmission, dissemination, and drift of norovirus strains in Europe (DIVINE-NET: http://www.eufoodborneviruses.co.uk/). In addition, sequence data have been used for the rapid detection of international norovirus outbreaks (e.g., caused by raspberries or on cruise ships) and the emergence of new global variants within the dominating genotype (*8,9*). In outbreak investigation, sequence data have been the most important tool used to link norovirus cases.

Viral genomes exist in different forms: DNA or RNA, single or double stranded, segmented or not segmented, plus or minus polarity, circular or linear. Viral genomes vary both in size (from 3,200 nucleotides to 1.2 million basepairs) and in complexity. Despite this diversity, a single method can be used to characterize and compare viral genes. Nucleotide sequencing provides the best differentiation between viral strains, and the polymerase chain reaction (PCR) is the favored technique for generating templates for nucleotide sequencing. The choice of PCR primers is crucial for generating the correct PCR product (gene or region) to be sequenced. Sequencing regions with high variability will usually yield the most appropriate information when closely related strains are to be compared, whereas more conserved regions should be sequenced when comparing distantly related strains. The genes coding for viral structural proteins are usually the most

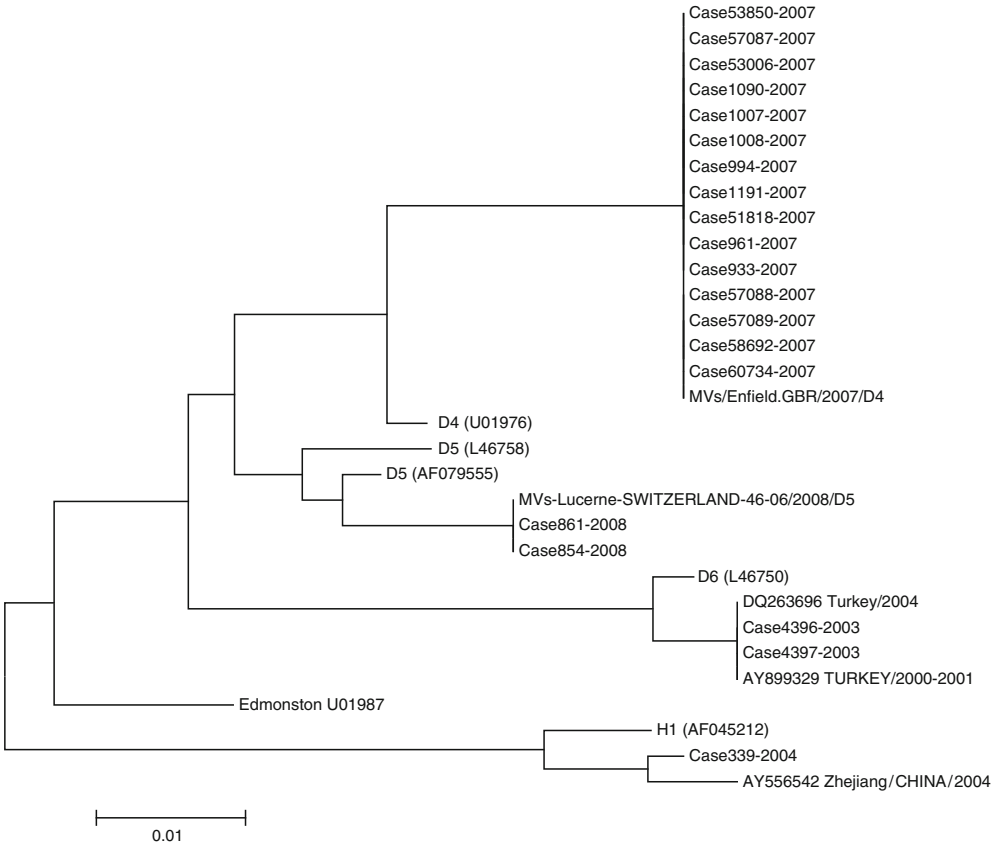


Fig. 1. Relationship between measles viruses from Norwegian cases and measles viruses detected in regions/countries the patients came from or had recently visited. The analysis was based on the N-gene sequence, and Norwegian cases were identical within each cluster. The Norwegian cases have been given names according to the year of isolation. The genotypes from Norwegian cases belonged to H1, D4, D5, and D6. The phylogenetic tree includes WHO Measles Reference Strains (2001) available on the EMBL database.

variable. Sequencing of variable genomic regions can be used to differentiate viruses into types, groups, subtypes, and strains, for example, human immunodeficiency virus (HIV) 1 and HIV-2; HIV-1 groups M, N, and O; the nine genetically distinct subtypes of HIV-1 group M; and various strains known as circulating or unique recombinant forms. It is also important to sequence regions that have been investigated in previous studies to have at disposition comparable data from a large number of other strains in nucleic acid databases.

The first step in nucleotide sequencing is template preparation by extraction of nucleic acids from samples likely to contain virus. There are many commercially available kits for manual viral nucleic acid extraction, as well as instruments that perform the extraction process automatically. In the case of DNA viruses the next step is PCR, whereas in the case of RNA viruses the genome has to be

reverse transcribed from RNA to complementary DNA (cDNA) before the PCR step.

DNA sequencing methods were developed in the mid-1970s. The two original DNA sequencing techniques are, however, very different in principle. In the enzymatic or dideoxy chain termination method of Sanger (*10*), a new DNA strand is synthesized from a template using a DNA polymerase, while in the chemical degradation method of Maxam and Gilbert (*11*) the original DNA is degraded. The dideoxy chain termination method is currently the most widely used technique for sequencing of viral genes and in the past decade has become an invaluable tool for molecular epidemiological investigations. In addition, the concepts of PCR technology have been utilized to enable the sequencing reaction to be cycled. A "cycled" dideoxy chain termination method, now known as cycle sequencing (*12*), forms the basis of sequencing reactions used in automated DNA sequencers. Automated methods have revolutionized the use of DNA sequences in molecular epidemiological investigations. Viral gene sequences are used to illustrate phylogenetic relationships visualized as "family trees" between viruses that elucidate viral evolution and possible routes of dissemination. The rapid evolution of viral genes is a significant advantage for studying relatedness and enables virus lineages to be differentiated even within an outbreak.

The purpose of this chapter is to provide a detailed protocol for the direct nucleic acid sequencing of PCR products generated from viral genes. Recommendations for protocols addressing the extraction of nucleic acids from sample material containing virus, reverse transcription of RNA genomes, and nucleic acid amplification by PCR are also included in this chapter. The method described is optimized for the ABI Prism® 310 Genetic Analyzer or ABI 3730 DNA Analyzer. While there are several other commercially available automated DNA sequencers, the biochemistry is common for all: dideoxy chain termination cycle sequencing using fluorescently labeled terminators.

2. Materials

2.1. Extraction of Viral Nucleic Acids

1. Any kits for isolation of viral nucleic acids may be used, and the kits include all required reagents and spin columns for isolation, for example, QIAamp Viral RNA Mini Kit for manual RNA isolation, QIAamp DNA Mini Kit for manual DNA isolation (Qiagen, Hilden, Germany) and MagNa Pure LC Total Nucleic Acid Isolation Kit for automated isolation of viral RNA or DNA (Roche, Mannheim, Germany).
2. MagNA Pure LC Instrument for automated nucleic acid isolation and equipment: tubes, trays, and pipets (Roche).

3. Microcentrifuge (e.g., Eppendorf centrifuge 5415D) for 0.5-mL microcentrifuge tubes.
4. Microcentrifuge tubes (e.g., Eppendorf polypropylene 0.5 mL with locking caps and Microamp PCR 0.2-mL tubes).
5. Automatic pipets capable of dispensing 0.5–20 μL and 10–100 μL.

2.2. Polymerase Chain Reaction and Template Preparation

The isolated nucleic acid can be used in PCR or reverse-transcriptase (RT)-PCR on different instruments/platforms and on standard thermal block cyclers.

1. Any kits for amplification of nucleic acids may be used, and the kits include all required reagents for amplification, such as, OneStep RT-PCR Kit for amplification of RNA viruses, Taq PCR Core Kit for amplification of DNA viruses, or QuantiTect Virus Kits for amplification of both RNA and DNA viruses (Qiagen).
2. Thermal cyclers (GeneAmp® 9700 or 2700, Perkin Elmer 2400, or Eppendorf Master Gradient) for cycle-sequencing reactions.
3. A temperature-cycling incubator capable of executing two consecutive programs over a temperature range of 45–95°C with an incubator chamber for 0.5-mL microcentrifuge tubes.
4. ExoSAP-IT® for PCR product cleanup; store at –20°C (USB Corp., OH). Eventually, use QIAquick-spin PCR purification kit (Qiagen).
5. Microcentrifuge (e.g., Eppendorf centrifuge 5415D) for 0.5-mL microcentrifuge tubes.
6. Microcentrifuge tubes (e.g., Eppendorf polypropylene 0.5 mL with locking caps and Microamp PCR 0.2-mL tubes).
7. Automatic pipets capable of dispensing 0.5–20 μL and 10–100 μL.

2.3. Cycle-Sequencing Reaction

1. Automatic sequencers with capillary electrophoresis technology: ABI Prism 310 Genetic Analyzer or Applied Biosystems 3730 DNA Analyzer (Applied Biosystems Inc., Foster City, CA) suitable for fluorescent sequencing using BigDye® Terminator chemistry and any primer.
2. Applied Biosystems 3730 DNA Analyzer equipment: 3130 and 3100 Series Plate Base 96-Wells, 3130 and 3100 series Plate Retainer 96-Wells, and plate Septa 96-Wells.
3. ABI Prism 310 Genetic Analyzer equipment: tubes and caps.
4. BigDye Terminator v1.1 Cycle Sequencing Kit, Applied Biosystems. The kit provides the required reagents for the sequencing reaction in a premixed, ready-to-use format. Single- or double-stranded DNA, PCR fragments, and large templates may be sequenced.
5. Sequencing primers. Any primer may be used. The PCR primer may also be used as the sequencing primer, although at a different concentration.

6. 3*M* sodium acetate, pH 4.6.
7. Ethanol, molecular biology grade and 70%.
8. 125 m*M* ethylenediaminetetraacetic acid (EDTA), pH 8.0.
9. Heater blocks or water baths capable of temperatures up to 95°C.
10. Hi-Di formamide, genetic analysis grade (Applied Biosystems).
11. POP7™ Performance Optimized Polymer and 10X buffer with EDTA (Applied Biosystems).

3. Methods

PCR is currently the favored technique for generating templates for sequencing of viral genes. This chapter therefore focuses on sequencing PCR amplicons. A few suggestions for isolation of viral RNA and DNA as well as preparation of PCR template are provided. For detailed procedures on nucleic acid purification and PCR template preparation, refer to manufacturers' protocols that are continuously improved and updated.

3.1. Extraction of Viral Nucleic Acids

Viruses contain either RNA or DNA genomes. There are several commercially available manual nucleic acid isolation kits for the isolation of viral RNA, viral DNA, or total viral nucleic acids. Many of the kits may be used for isolation of viral nucleic acids from a variety of specimens (serum, plasma, whole blood, urine, cerebrospinal fluid, cell culture supernatant, stool, tissue, cell-free body fluid). There are also several kits for isolation of viral nucleic acids with automated nucleic acid extractors (e.g., ManNA Pure, Roche). Any manual or automated protocol should work well. In our laboratory we use Qiagen kits (e.g., QIAamp Viral RNA Mini Kit and QIAamp DNA Mini Kit) for manual nucleic acid isolation and MagNA Pure LC Total Nucleic Acid Isolation Kit for automated nucleic acid isolation from clinical specimens. These kits provide nucleic acid free of PCR inhibitors.

3.2. PCR Template Preparation

There are several kits for PCR amplification of isolated nucleic acids, and the PCR amplicons are used as templates for the sequencing reaction. The isolated DNA from DNA viruses may be amplified directly (e.g., Taq PCR Core Kit, Qiagen), whereas isolated RNA from RNA viruses must be reverse transcribed to cDNA before the PCR amplification step. One-step RT-PCR kits (e.g., OneStep RT-PCR Kit, Qiagen) provide a fast and successful alternative to performing a separate reverse-transcriptase and PCR reaction. Any primers may be used with these kits (specific primers, random primers, etc.).

3.3. PCR Product (Template) Purification

The most important factor for obtaining good sequence data with dye terminators is a clean, unique target with a single binding site for the primer (*see* **Notes 1–4**). There are many protocols for purification of PCR products prior to sequencing, and any protocol to remove deoxynucleotide 5′-triphosphates (dNTPs) and primers should work.

ExoSAP-IT (shrimp alkaline phosphatase [SAP] and exonuclease I [Exo I]) purification is a simple, fast, and efficient method for the purification of PCR products for sequencing. The ExoSAP-IT degrades nucleotides and single-stranded DNA (primers) and is particularly useful when limiting concentrations of primers and nucleotides cannot be used in PCR.

1. Set up a reaction with ExoSAP-IT (one tube for each reaction/template) by mixing 7 μL of PCR product and 2 μL of ExoSAP-IT. Vortex briefly.
2. Run on thermocycler as follows: One cycle at 37°C for 15 min, 80°C for 15 min, and 4°C indefinitely.
3. The samples may be kept at 4–8°C until cycle sequencing is performed. Long-term storage should be at –20°C.
4. PCR product purification with QIAquick PCR Purification kit (Qiagen) is also compatible with BigDye chemistry. This method is suitable for PCR fragments ranging from 100 to 1,000 bp. See the manufacturer's manual for a detailed protocol.

3.4. Cycle-Sequencing Reaction

Cycle sequencing using BigDye Terminator chemistry provides reproducible results for sequencing of PCR fragments (*see* **Notes 1–5**). The method is quick, convenient, and commonly used. It is important to sequence the template in both directions to minimize sequencing errors and to include a control DNA template in each set of sequencing reactions. Control DNA is included in the kit.

The BigDye Terminator protocol recommends the use of 8 μL of the Big Dye v1.1 Terminator Ready Reaction mix in a final reaction volume of 20 μL. Good results are, however, obtained with half of the recommended volume when purified PCR fragments are sequenced.

1. For each reaction (total volume 10 μL), add the following reagents to a separate tube:
   a. 4 μL Big Dye v1.1 Terminator Ready Reaction mix.
   b. 1 μL 3.2 μ*M* primer (final concentration 3.2 pmol).
   c. 1–5 μL template (2–10 ng purified PCR product/μL).
   d. 0–4 μL distilled H_2O.
2. Mix well, spin briefly, and run on thermocycler as follows:

   Step 1: 96°C 1 min.

   Step 2: 96°C 10 s.

Step 3: 50°C 5 s.

Step 4: 60°C 1 min.

Step 5: Repeat steps 2–4, 25 times.

Step 6: 4°C, indefinitely.

3. Hold at 4°C until ready to purify.

3.5. Purification of Extension Products

The unincorporated dye terminators must be removed prior to capillary electrophoresis (*see* **Note 6**). The ethanol/EDTA/sodium acetate precipitation protocol works well and is known to be suitable for generating clean sequences using BigDye Terminator v1.1 Cycle Sequencing Kits. Other methods may also be used.

1. For each sequencing reaction, prepare a separate 1.5-mL microcentrifuge tube with the following:
   a. 1 µL 3*M* sodium acetate, pH 4.6.
   b. 25 µL 96% ethanol.
   c. 1 µL 125 m*M* EDTA, pH 8.0.
   d. 10 µL sequencing product.
2. Mix briefly.
3. Keep on ice for 10 min.
4. Spin in a microcentrifuge at maximum speed for 15–30 min at 4°C.
5. Carefully remove the supernatant completely with a pipet and discard.
6. Add 80 µL 70% ethanol to the pellet.
7. Keep at room temperature for 2–3 min (may be stored overnight at –20°C).
8. Remove the supernatant completely.
9. Dry the pellet on a thermocycler (or heat block) with caps open for 3 min at 60°C. It is important that the pellet is completely dried. The dried pellet may be stored at –20°C until being prepared for automatic sequencing.
10. Dissolve the dried pellet in 15 µL Hi-Di formamide immediately before preparing the samples for automatic sequencing. Samples with Hi-Di formamide should not be kept at room temperature for more than 2–3 d. Long-term storage should be at –20°C.
11. Heat the tubes with closed caps for 2 min at 94°C on a thermocycler (heat block).
12. Remove the tubes immediately and place on ice for 1–2 min.
13. Spin the tubes in a microcentrifuge at maximum speed for 2 s.
14. Load the samples for capillary gel electrophoresis as follows:

a. ABI Prism 310 Genetic Analyzer: Load 15 μL of reconstituted sample to the tubes belonging to the instrument.

b. ABI 3730 DNA Analyzer: Dilute each reconstituted sample 10X in Hi-Di formamide. Load 10 μL of this 1:10 dilution to microtiter plate wells belonging to the instrument.

15. All of the empty wells in the microtiter plate must be filled with Hi-Di formamide.

3.6. Capillary Gel Electrophoresis and Data Collection

Prepare the automatic sequencer according to the manual provided with the equipment, and use POP7™ polymer and 10X buffer with EDTA from Applied Biosystems to fill the capillaries. After preparation, load the samples (tubes or tray), set up the data collection software, and start the sequencer according to the manual. Make sure that the thin capillary is aligned properly with the sample in the tray/tube to take up the entire DNA sample to be sequenced. The sequencing products are subsequently separated and read by the automatic sequencers (*see* **Notes 6–8**).

The BigDye Terminator chemistry uses four different fluorescent labels, allowing the reaction to be analyzed in a single lane. Briefly, the fluorescently labeled DNA fragments (the terminated extension products) are separated and identified using capillary gel electrophoresis. A laser at the end of the capillary excites the ddNTP (dedioxynucleotides) dyes, causing the incorporated labeled dideoxynucleotides to illuminate different colors. The colors are analyzed by the computer, and the user is provided with a chromatogram and the suggested DNA sequence.

3.7. Sequence Analysis and Assessment

Automatic sequencers compile and deliver a computer file with a chromatogram and a suggested sequence. In most cases the chromatograms and sequences delivered by the sequencer should be carefully interpreted and manually edited before using the sequence data. The automatic analysis of the chromatogram peaks may be incorrect, especially early in the sequence (the first 40–50 bases from the primer binding sites) or at the end when the resolution of large fragments is not optimal (*see* **Notes 7–9**). There are several programs for comparing and editing sequences (e.g., Sequencher 4.5 and Bioedit). The edited sequences should then be compared with sequences from different databases (EMBL or GenBank) to gain as much information about the sequences as possible. The databases are available through the Internet, and information may be obtained using different search tools (e.g., FASTA and BLAST).

3.8. Phylogenetic Analysis

Phylogenetic analysis provides important molecular epidemiological information. The aim of phylogenetic analysis is to compare sequences, analyze gene families, and estimate evolutionary relationships.

The result of molecular phylogenetic analysis is visualized as a phylogenetic "tree" or dendrogram. DNA or RNA of closely related organisms usually exhibits a high degree of sequence similarity. Molecular phylogeny uses such sequence heterogeneity to build a "relationship tree" that illustrates the probable evolution of various organisms. There are numerous software programs that estimate phylogenies, and many are free of charge via the Internet.

The quality of phylogenetic analysis largely depends on the quality of the sequence alignment. Interactive software programs generate both multiple-sequence alignments (e.g., ClustalW, Pileup) and phylogenetic trees. The MEGA (Molecular Evolutionary Genetics Analysis) 4.5 software is a user-friendly interactive program that constructs multiple alignments and phylogenetic trees. For example, when using MEGA to generate a phylogenetic tree, sequences are automatically aligned using ClustalW.

4. Notes

1. Template concentration and purity are the two most common causes of poor or no sequence data. The amount of template used in a sequencing reaction can affect the quality of sequence data, and there is a "threshold" amount that must be used to generate any sequence data. The recommended amount of PCR product is 1–3 ng for 100- to 200-bp fragments, 3–10 ng for 200- to 500-bp fragments, and 2–20 ng for 500- to 1,000-bp fragments *(13)*. In general, the optimal concentration of template may be determined by multiplying the length of the template in kilobases by 25 ng. Too much DNA may cause premature termination of signal. This occurs when the dNTPs in the cycle-sequencing reaction are distributed among too many extending chains. The dNTPs will be depleted early in the reaction and thus yield an excessive amount of short fragments. Quantification of template may be determined by gel electrophoresis. Fluorescent sequencing is very sensitive to certain contaminants in the DNA sample, including dNTPs, primers, and salts. It is critical to remove excess PCR primers from the sequencing reaction. PCR primers will act as sequencing primers and lead to extra bands that correspond to the complementary strands from opposite orientations. Thus, incomplete removal of PCR primers prior to sequencing may yield ambiguous results that are visualized as sequences with numerous double peaks at single positions. Excess dNTPs will disturb the specific ratios of dNTPs/ddNTPs in the sequencing reaction.
2. It is imperative to ensure that the PCR product to be sequenced is the correct fragment. Multiple PCR products in a single

sequencing reaction will yield ambiguous sequences. Visualizing PCR products on an agarose gel will give a good indication of the quality of the product. In the case of multiple products (bands), gel purification of the desired product is necessary. Gel separation of PCR products may, however, be difficult if the products are similar in size (e.g., amplifying related DNA). In this case, optimization of the PCR reaction may be necessary, or new PCR primers may need to be designed to use a more specific priming site. Restriction sites in PCR fragments may also be used to identify the correct product bands. Alternatively, nested primers may be employed to reamplify the desired product. Nested primers will verify the identity of the product and simultaneously eliminate any unwanted products.

3. PCR primers that are used as sequencing primers must be suitable for the cycle sequencing conditions. While inefficient primers are sometimes acceptable for PCR, the same primers may fail in sequencing (which is a linear amplification). The melting temperature T_m of sequencing primers should be between 50 and 60°C. In addition, the primers must not form primer-dimers as this will deplete the availability of primers needed for the sequencing reaction. Sequencing primers should be 18–24 nucleotides in length and approximately 50% in GC content *(14)*. High GC content or long primers may increase the formation of secondary structures that influence the melting temperature. It is preferable to choose primers with a low melting temperature.
4. Removal of components that may inhibit the sequencing reaction is necessary. Nuclease contamination in a template preparation, as well as repeatedly thawing/freezing samples, can degrade DNA over time. High concentrations of impure DNA may also contain a larger proportion of contaminants (excess primers, dNTPs, salts) that may reduce the quality of the DNA sequence generated. Generally, reisolation and purification of the template DNA are necessary to obtain good DNA sequences. It is wise to limit the time and intensity of UV illumination to a minimum when extracting PCR products from gels to reduce DNA degradation. Prepare fresh stocks of commonly used reagents, such as buffers, using high-quality distilled water.
5. Several factors may result in early termination of sequence data throughout the sequencing reaction, including template concentration, deoxyribonuclease (DNase) contamination, and secondary structures. Secondary structures that do not melt during cycle sequencing can cause premature termination of sequences. Addition of DNA denaturants (e.g., formamide or dimethyl sulfoxide [DMSO]) to the sequencing reaction may

reduce early termination. Denaturants may melt duplex formation and enable optimal polymerase activity. Changing the cycle-sequencing parameters to include a higher denaturation temperature (98 vs. 96°C) and eliminating the 50°C annealing step may be useful. The 60°C cycle will in this case function as both the annealing and extension steps. High salt concentrations may also result in premature termination. Sequencing the opposite strand will sometimes yield better results.

6. Purification of sequencing reactions is important to gain good sequence data. Purification using ethanol/EDTA/sodium acetate precipitation is recommended when a good signal from the first base is required (*see* the BigDye kit manual). However, it is important to use the correct ethanol concentration ***(14)***. Too high ethanol concentrations will result in precipitation of residual terminators along with the sequencing products, whereas too low concentrations will result in no signal due to a failed reaction. Ethanol precipitation also removes excess salts. Ethanol contamination may also occur when the sample is insufficiently dried after precipitation and may inhibit sequencing reactions.
7. Multiple peaks under the primary sequence peak and many "N"s within the sequence may indicate the presence of two nucleotides at the same position (polymorphisms), high background, or the presence of multiple products. High background may also be a problem if contaminated reagents are used for either template preparation or sequencing reactions. When background poses a problem, it is necessary to view the average signal strength and edit the sequences manually.
8. Low signal strength may be the result of too little or degraded DNA or primer, inhibitory components, contaminated reagents, or poor primer binding. A correct primer concentration and annealing temperature are critical.
9. Difficulties with the sequencer may be due to improper capillary filling when fresh polymer is being pumped through the array. See the manual for guidelines addressing instrument-related problems.

References

1. Powell, S. C., and Attwell, R. (1999). The use of epidemiological data in the control of foodborne viruses. *Rev. Environ. Health* **14**, 31–37.
2. Holmes, E. C., Nee, S., Rambaut, A., Garnett, G. P., and Harvey, P. H. (1995). Revealing the history of infectious disease epidemics through phylogenetic trees. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **349**, 33–40.
3. Suss, J., Beziat, P., and Schrader, C. (1997). Viral zoonosis from the viewpoint of their epidemiological surveillance: tick-borne encephalitis as a model. *Arch. Virol.* **13**, (Suppl.) 229–243.

4. Hu, D. J., Dondero, T. J., Rayfield, M. A., George, J. R., Schochetman, G., Jaffe, H. W., et al. (1996). The emerging genetic diversity of HIV. The importance of global surveillance for diagnostics, research, and prevention. *JAMA* **275**, 210–216.
5. Løvoll, Ø., Vainio, K., and Skutlaberg, D. H. (2004). Measles outbreak in Norway in children adopted from China. *Euro Surveill.* **8**, 21.
6. Løvoll, Ø., Vonen, L., Nordbo, S. A., Vevatne, T., Sagvik, E., Vainio, K., et al. (2007). Outbreak of measles among Irish travellers in Norway: an update. *Euro Surveill.* **12**, E070614.2
7. Schmid, D., Holzmann, H., Abele, S., Kasper, S., König, S., Meusburger, S., et al. (2008). An ongoing multi-state outbreak of measles linked to non-immune anthroposophic communities in Austria, Germany, and Norway, March–April 2008. *Euro Surveill.* **13**, 16.
8. Lopman, B., Vennema, H., Kohli, E., Pothier, P., Sanchez, A., Negredo, A., et al. (2004). Increase in viral gastroenteritis outbreaks in Europe and epidemic spread of new norovirus variant. *Lancet* **363**, 682–688.
9. Siebenga, J. J., Vennema, H., Duizer, E., and Koopmans, M. P. (2007). Gastroenteritis caused by norovirus GGII.4, The Netherlands, 1944–2005. *Emerg. Infect. Dis.* **13**, 144–146.
10. Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U S A* **74**, 5463–5467.
11. Maxam, A. M., and Gilbert, W. (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U S A* **74**, 560–564.
12. Murray, V. (1989). Improved double-stranded DNA sequencing using the linear polymerase chain reaction. *Nucleic Acids Res.* **17**, 8889.
13. BigDye® Terminator v1.1 Cycle Sequencing Kit Protocol. Part Number 4337036 Rev. A 09/2002. Applied Biosystems.
14. Roswell Park Cancer Institute. Sequencing Basics. Available at: http://www.roswellpark.org/Research/Shared_Resources/Biopolymer_Resource/DNA_Sequencing/Sequencing_Basics. Accessed January **24**, 2008.

Chapter 17

Full Sequencing of Viral Genomes: Practical Strategies Used for the Amplification and Characterization of Foot-and-Mouth Disease Virus

Eleanor M. Cottam, Jemma Wadsworth, Nick J. Knowles, and Donald P. King

Abstract

Nucleic acid sequencing is now commonplace in most research and diagnostic virology laboratories. The data generated can be used to compare novel strains with other viruses and allow the genetic basis of important phenotypic characteristics, such as antigenic determinants, to be elucidated. Furthermore, virus sequence data can also be used to address more fundamental questions relating to the evolution of viruses. Recent advances in laboratory methodologies allow rapid sequencing of virus genomes. For the first time, this opens up the potential for using genome sequencing to reconstruct virus transmission trees with extremely high resolution and to quickly reveal and identify the origin of unresolved transmission events within discrete infection clusters. Using foot-and-mouth disease virus as an example, this chapter describes strategies that can be successfully used to amplify and sequence the full genomes of RNA viruses. Practical considerations for protocol design and optimization are discussed, with particular emphasis on the software programs used to assemble large contigs and analyze the sequence data for high-resolution epidemiology.

Key words: Complete genome, foot-and-mouth disease virus, nucleotide sequence, virus.

1. Introduction

During the past 15 years, a number of incremental improvements have been made to methods used to generate nucleotide sequence data. The principle underpinning the mostly widely used sequencing approaches is based on the dideoxynucleotide

D.A. Caugant (ed.), *Molecular Epidemiology of Microorganisms*, Methods in Molecular Biology, Vol. 551
DOI: 10.1007/978-1-60327-999-4_17,

chain-termination method initially devised by Fred Sanger in the 1970s (*1*). The throughput and robustness of these methods have been improved by the use of fluorescent dyes and capillary separation technologies, such that the routine assembly of large fragments of genomic DNA (>10 kb) is now achievable by many modestly equipped laboratories. For the large part, protocols developed to sequence large fragments of nucleic acid can also be adapted to characterize the genomes of RNA viruses, which typically are 15 kb or less. Full-genome sequences of viruses can be used to address fundamental questions relating to evolution, identification of critical antigenic determinants, and viral molecular epidemiology. Although sequencing small numbers of some viral genomes can be straightforward, specific protocols and work flows are required to effectively manage projects that aim to characterize the molecular epidemiology of viral transmission.

Using foot-and-mouth disease virus (FMDV) as an example, this chapter describes strategies that can be successfully used to amplify and sequence the complete genomes of RNA viruses. Foot-and-mouth disease (FMD) is a highly contagious disease affecting cloven-hoofed livestock (cattle, sheep, pigs, goats, and water buffalo). The causative agent is a virus belonging to the genus *Aphthovirus* (family: Picornaviridae) that exists as seven antigenically distinct serotypes, each comprising numerous and constantly evolving variants (*2*). The genome of FMDV is approximately 8,300 nucleotides in length. It comprises a polyadenylated positive-sense RNA that encodes a single polyprotein, which is posttranslationally cleaved into constituent capsid proteins and nonstructural proteins involved in viral replication.

In common with most other RNA viruses, the enzyme (RNA-dependent RNA polymerase) responsible for replication of the FMDV genome has poor fidelity, such that changes to the nucleotide sequence frequently occur and are inherited to progeny viruses. This rapid evolution rate of FMDV allows virus transmission trees to be reconstructed with extremely high resolution, opening up the possibility of using these data to retrospectively reveal and identify the origin of unresolved transmission events (*3,4*). In addition to forensic molecular epidemiology, full-genome sequence data have also recently contributed to our understanding of a number of aspects of FMDV evolution, including (i) evolutionary rates (5); (ii) sites and importance of recombination (*6,7*); (iii) identification of ordered RNA structures (*8*); and (iv) contribution and significance of the quasi-species phenomenon to evolution (*9*). Sequence data from a wide variety of FMDV isolates also play an important role in the reiterative design of oligonucleotide primers used for molecular assays for routine diagnostic use in reference laboratories (for pan-reactive and serotype-specific detection and strain characterization).

1.1. Amplification Strategies: Design and Targeting of Polymerase Chain Reaction Primers

The extent of the run length obtained by capillary sequencers places a limit on the maximum distance between oligonucleotide primers (either in the polymerase chain reaction [PCR] amplification or cycle sequencing setup stages). In contrast to DNA targets, which are relatively stable, researchers who study RNA viruses, such as picornaviruses, are familiar with the plasticity of viral genomes. This high variability poses particular challenges for the design of pan-reactive oligonucleotide primers to reliably amplify complete viral complementary DNAs (cDNAs). For viruses such as FMDV, the existence of multiple serotypes (whose nucleotide sequences may vary by as much as 50% in some genome regions) can further complicate the identification of suitable target sequences.

As a consequence, the two extremes of the sequencing strategies used for FMDV are illustrated in **Fig. 1** (*see* **Fig. 1a,b**) and shown by representative agarose gels in **Fig. 2**. In both of these

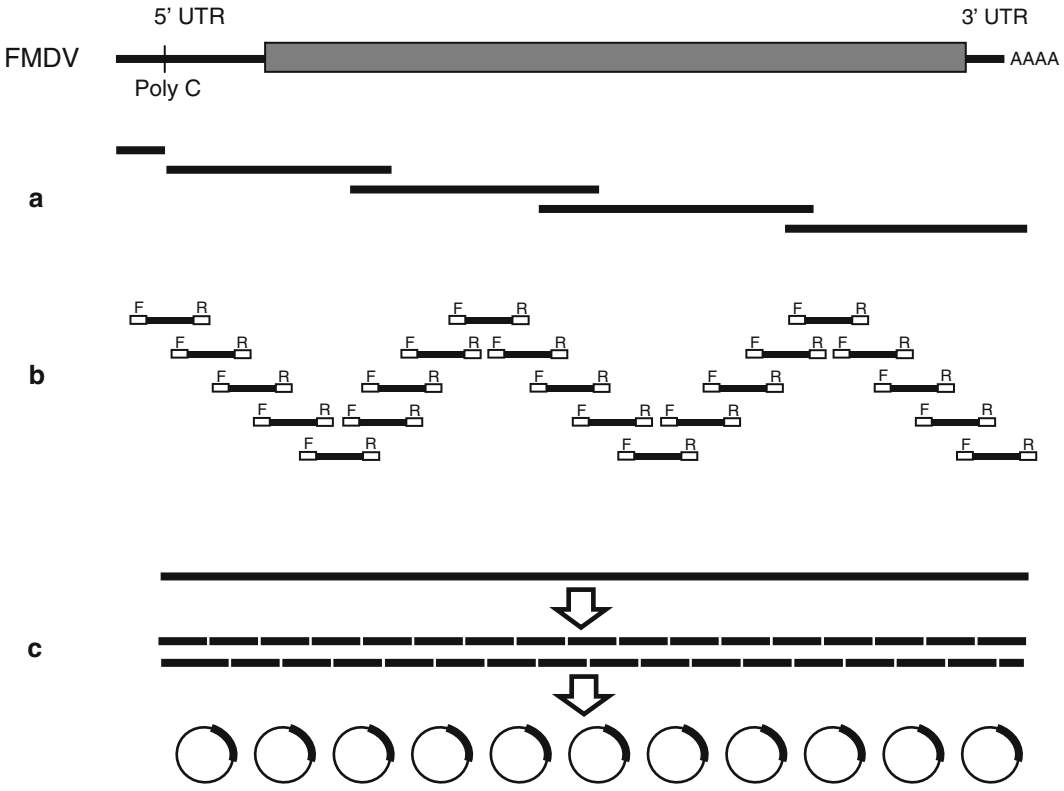


Fig. 1. Outline of RT-PCR strategies that have been used to amplify the complete genome sequences of FMDV. (**a**) Long overlapping products (~3 kb) are generated by PCR using full-length cDNA as a template. Sequences are obtained using a panel of specific sequencing primers (*see* **ref. 3**). (**b**) Short products (~700 bp) are generated using FMDV-specific primers, which also incorporate regions (labelled F and R) targeted by the sequencing primers (*10*). (**c**) Long-range RT-PCR is used to amplify a product comprising the complete L fragment of FMDV. This may either be sequenced using many specific primers (*11*) or can be fragmented by restriction digest and cloned into a bacterial plasmid vector (pilot studies using this method have been undertaken by IAH in collaboration with the Wellcome Trust Sanger Institute, Cambridge).

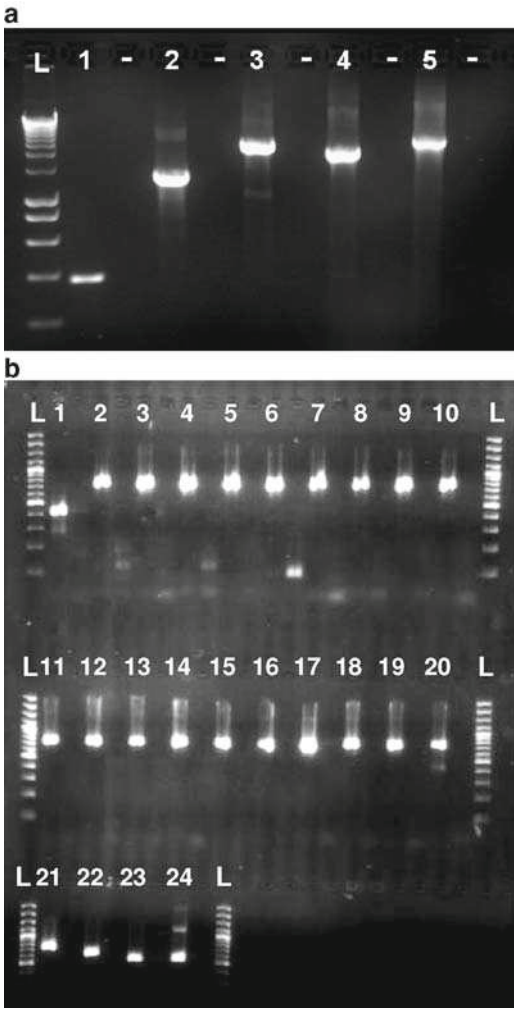


Fig. 2. Amplification of full-FMDV genomes by RT-PCR. Agarose gel electrophoresis shows RT-PCR fragments representing amplification of the entire FMDV genome. (**a**) amplification of O/UKG/2001 genome using 5 RT-PCR fragments and (**b**) amplification of O/UKG/2007 genomes using 24 RT-PCR fragments.

approaches (adopted for the characterization of FMD outbreaks in the United Kingdom in 2001 and 2007), a large number of specific primers were required. Furthermore, these oligonucleotides are specific for defined lineages of FMDV, limiting their use for study of other genotypes of FMDV. Other FMD laboratories have used similar approaches also requiring large numbers of primers (*11–15*). Additional protocols, such as rapid amplification of cDNA ends (RACE), can be used to generate sequence data for the terminal ends of the genome and regions close to the poly (C) tract of FMDV. To reduce the complexity of the cycle-sequencing reactions, recognition sequences for universal sequencing primers (such as M13) can be incorporated into the 5′ ends of the primers used for PCR (*see* **Fig. 1a**).

Alternative approaches, such as shotgun cloning (for example, **Fig. 1c**) are also being considered for full-genome sequencing. Initially, these use long-range PCR to amplify large fragments of the virus genome (possibly even encompassing entire genomic sequences). These PCR products are subsequently fragmented and cloned into plasmid vectors prior to sequencing and reconstruction of the viral sequence. Since this approach uses only two viral-specific primers (which can be targeted to highly conserved regions) and is not reliant on internal virus-specific primers, this method may provide a more suitable approach that has a broader sensitivity to different viral variants. However, these methods need to balance the advantages in diagnostic sensitivity that are gained from using a smaller number of primers with the drawback of lower analytical sensitivity that may arise from amplifying large PCR products (in comparison to shorter fragments).

1.2. Overview of Approaches Used for FMDV Sequencing

In this chapter, a guide protocol that has been successfully used to sequence FMDV is presented. Although some of the finer details are specific to FMDV, the general approaches described are broadly applicable to other RNA viruses. Indeed, similar methods have been described recently to characterize the genomes of other viruses that infect humans, livestock, and plants (*16–23*).

2. Materials

2.1. RNA Extraction

1. 0.04*M* phosphate-buffered saline: 35 m*M* Na_2HPO_4, 5.7 m*M* KH_2PO_4, pH 7.6. Store at room temperature.
2. Sterile sand (Fine Sifted, BDH). Small aliquots (~3 g) are prepared and autoclaved prior to use. Store at room temperature.
3. Sterile pestle and mortar (Fisher); autoclave prior to use.
4. TRIzol Reagent (Invitrogen). Store at +2 to 8°C. This solution contains phenol and guanidine isothiocyanate; care should be taken to minimize skin contact and inhalation.
5. Chloroform (AnalaR Grade, BDH) (toxic and probable carcinogen; care should be taken to minimize inhalation and ingestion).
6. 0.2*M* glycogen (Roche).
7. Isopropanol (propan-2-ol) (AnalaR Grade, BDH).
8. Ethanol (AnalaR Grade, BDH). Store at +2 to 8°C.
9. Nuclease-free water (deoxyribonuclease [DNase] and ribonuclease [RNase] free) (Invitrogen). Store at room temperature.

2.2. Reverse Transcription and PCR Amplification

1. Random hexamers (Promega). Store at –20°C.
2. Deoxynucleotide 5′-triphosphate (dNTP) mixture (Promega). The dNTP mix is a premixed solution containing sodium salts of dATP (deoxyadenosine 5′-triphosphate), dCTP (deoxycytosine 5′-triphosphate), dGTP (deoxyguanosine 5′-triphosphate), and dTTP (deoxythymidine 5′-triphosphate), each at 10 m*M* in water. Store at –20°C.
3. Oligonucleotide primers (Sigma-Aldrich). Complete list of primers used for PCR amplification of FMDV are described elsewhere (***2,3,10***).
4. Reverse transcription kit: SuperScript™ III RT (Invitrogen). Enzyme is supplied with a vial of 5X first-strand buffer (250 m*M* Tris-HCl, pH 8.3, 375 m*M* KCl, 15 m*M* $MgCl_2$) and a vial of 100 m*M* dithiothreitol (DTT). Store at –20°C.
5. RNaseOUT (Invitrogen); store at –20°C.
6. GFX™ PCR DNA and Gel Band Purification Kit (GE Healthcare).

2.3. PCR Cleanup Prior to Setting Up Sequencing Reactions

1. Agarose (UltraPure™, Invitrogen). Store at room temperature.
2. Tris-borate-ethylenediaminetetraacetic acid (EDTA) buffer (National Diagnostics). 10X solution: When diluted, the 1X solution contains 89 m*M* Tris-base, 89 m*M* boric acid (pH 8.3), and 2 m*M* Na_2EDTA. Store at room temperature.
3. Ethidium bromide (UltraPure). Of a 10 mg/mL stock solution, added 2 μL to 100 mL gels to visualize PCR bands. Ethidium bromide is a potent mutagen. Therefore, care should be taken to minimize exposure and to ensure correct disposal of material (solutions and gels) containing ethidium bromide. Store at room temperature.
4. 6X loading buffer for samples to be tested by agarose gel electrophoresis (Invitrogen).
5. DNA standards, if required (Invitrogen).

3. Methods

3.1. RNA Extraction (see Notes 1 and 2)

1. Using sterile sand and a pestle and mortar, prepare a 10% (w/v) suspension of the tissue sample in phosphate-buffered saline. Liquid samples (such as serum) can be processed straight to step 3. Depending on application and nature of the sample to be tested (*see* **Note 3**), alternative RNA extraction protocols can also be used (such as commercially available silica-based spin columns).
2. Centrifuge at 300g for 10 min.

3. Add 200 µL of the sample supernatant to 1 mL of TRIzol reagent in a microfuge tube (*see* **Note 4**).
4. Add 240 µL of chloroform directly to the tube.
5. Mix the tube by inversion and centrifuge for 15 min at 10,000g at 2–8°C.
6. Transfer the top phase to a fresh microfuge tube and add 1µL of 0.2M glycogen.
7. Add an equivalent volume of isopropanol.
8. Mix the tube by inversion and centrifuge for 15 min at 10,000g at 2–8 °C.
9. Carefully wash the pellet (containing the RNA) with ice-cold 70% ethanol and recentrifuge for 15 min (10,000g at 2–8 °C).
10. Air-dry the pellet and resuspend RNA in nuclease-free water.

3.2. Reverse Transcription and PCR Amplification

1. Prepare primer mix (9 µL) containing 30 pmol reverse primer [5′-GGC GGC CGC TTT TTT TTT TTT TTT-3′], 50 ng random hexamers, and 30 nmol of each dNTP (3 µL of a 10 m*M* solution).
2. Add to 12 µL prepared RNA.
3. Denature the RNA by incubating the RNA/primer mix at 70°C for 3 min and place on ice for 3 min.
4. Add 17 µL of reverse-transcription (RT) mix containing 8 µL first-strand buffer, 2 µL 0.1 mM DTT, 2 µL RNaseOUT, 5 µL nuclease-free water.
5. Add 2 µL Superscript III Reverse Transcriptase.
6. Incubate at 42°C for 1–4 h followed by 85°C for 5 min. A specific PCR amplifying the 5′ end of the genome can be used to test that complete first-strand cDNAs have been generated.
7. Cleanup cDNA using GFX PCR DNA and Gel Band Purification kit according to manufacturers' instructions and elute in 50 µL. This step removes unincorporated primers and dNTPs from the RT reaction.
8. Set up a PCR master mix in a clean room using the primer sets required for amplification of the genomic fragments.
9. Add 2.5 µL cDNA to each reaction in a separate area away from the PCR clean room (*see* **Note 5**).
10. Run thermocycling program (as described in refs. **2**, **3**, **10**; *see* **Note 6**).

3.3. PCR Cleanup Prior to Setting Up Sequencing Reactions

1. Run 2 µL of PCR product on 1.2% (w/v) agarose gel at 105 V for 30 min to check reaction has worked.
2. Clean up cDNA using GFX PCR DNA and Gel Band Purification kit according to manufacturers' instructions.

3. Quantify DNA concentration in purified PCR product. This can be done using a spectrophotometer (e.g., Nanodrop, Thermo Fisher Scientific) or by agarose gel electrophoresis using DNA standards (*see* **Subheading 2.3**.).
4. Dilute products to give appropriate concentrations for sequencing.
5. Prepare sequencing reaction using diluted PCR product.

3.4. Analysis of Sequence Data

Sequencing viral genomes can quickly accumulate a large amount of data (*see* **Note 7**). Software programs (such as Lasergene, http://www.dnastar.com/) can be used to simplify the alignment of individual sequences and to rapidly assemble large contigs. The minimum criterion for acceptance of a final sequence is that each nucleotide position should be determined by sequencing reactions in either direction (forward and reverse).

Currently, the genetic evolution and relationships of viruses are studied by analyzing their genetic sequence data by phylogenetic methods. Phylogenetic trees are constructed and used to deduce the genetic relatedness of the viruses. There are different methods for constructing phylogenetic trees; the first approach developed was the maximum parsimony methodology, but more recently maximum likelihood (*24*) and Bayesian methods (*25*) are the preferred techniques for tree construction. Other methods based on distance matrixes, such as neighbor-joining (*26*) or unweighted pair-group method with arithmetic mean (UPGMA) (*27*), which calculate genetic distance from multiple sequence alignments, are simpler to implement but do not invoke an evolutionary model.

Maximum parsimony determines the most parsimonious tree requiring the least evolutionary steps. This method is simple and as such makes very few assumptions about the evolutionary process. However, certain features of genetic evolution of organisms present problems when using this method of tree construction. First, inaccuracies can occur as a result of the existence of homoplasy. Homoplasy describes processes, such as convergent evolution, by which a single mutation can occur twice on independent branches of a tree. Hence, it implies that two sequences sharing a mutation were not necessarily derived from a common ancestor that also contained this mutation. Another hurdle to overcome is back-mutation, by which a mutation reverts to its original genotype. This can cause the specific sequence to appear more ancestral than is necessarily the case. A further drawback to the method of maximum parsimony is that it takes no account of the rate at which mutations arise and the varying probabilities of different mutations occurring (i.e., transversions vs. transitions).

For these reasons, the parametric method of maximum likelihood is usually preferred as it provides the most probable tree that suits a specific determined evolutionary model. Providing that the model employed is a reasonable approximation of the evolutionary processes that gave rise to the observed genetic data,

this analysis is potentially more powerful than other methods. The evolutionary model may include a large number of parameters accounting for differences in the probabilities of various character states, differences in the occurrence of particular substitutions/mutations, and differences in the probabilities of change among characters. With the sophisticated models such as the Hasegawa-Kishino-Yano (HKY) model (*28*) and the general time reversible (GTR) model (*29*), an improved idea of phylogeny is achieved, although fitting an incorrect model can give incorrect results. The suitability of models can be tested using a program such as model test (*30*).

Maximum likelihood estimation of tree phylogeny is generally preferable to maximum parsimony because it is statistically consistent with a better statistical foundation, and it allows complex modeling of evolutionary processes. However, the maximum likelihood method has a computing limitation for large numbers of sequences. To infer statistical confidence in either maximum parsimony or maximum likelihood, constructed phylogenies bootstrap analyses (*31*) are performed. A further method to infer phylogenies is that of Bayesian inference, which generates a posterior distribution for a parameter based on the prior for that parameter and the likelihood of the data (represented by the sequence alignment). In other words, whereas maximum likelihood analysis investigates the probability of the observed data given a specific evolutionary model, Bayesian inference looks at the probability that a model is correct given the observed data set. With the availability of Markov chain Monte Carlo methods (*32*), Bayesian inference can be a preferred choice for tree estimation because it can be faster than maximum likelihood, and no bootstrapping is required as the posterior probabilities determine the statistical confidence in the tree.

Although in the majority of incidences maximum likelihood or Bayesian inference is preferable for tree construction, in certain situations maximum parsimony can be a viable alternative. When studying closely related sequences over a short time period the likelihood of back-mutation is relatively low, and hence maximum parsimony tree construction is likely to give an accurate estimation of tree phylogeny. Phylogenetic analysis of virus sequences is often performed with the aim of tracing specific virus history, and in these cases the method of statistical parsimony can be used. The distances depicted by parsimony trees represent the actual number of differences between sequences, whereas for a maximum likelihood tree the probability of change is shown (**Fig. 3**).

Often when studying viruses, closely related sequences are being investigated, with a focus on the accumulation of changes, and in this case a simpler representation of the raw data as depicted by parsimony is desirable. The TCS statistical parsimony program (*34*) can position sequences internally on a branch, which assists in

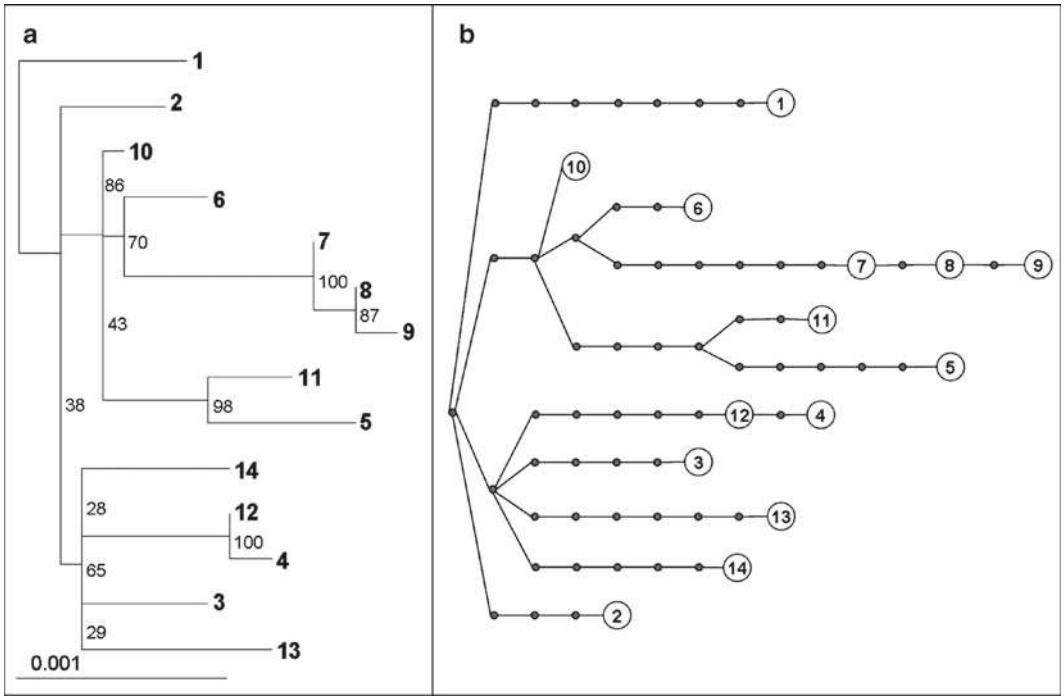


Fig. 3. Phylogenetic analysis of FMDV genomes from the 2001 outbreak in the United Kingdom. (**a**) Maximum likelihood phylogenetic tree representing 14 FMDV complete genomes rooted to sequence "1," constructed using PhyloWIN95 (*33*), incorporating the HKY model of nucleotide substitution with gamma distributed rate heterogeneity. Bootstrap values from 1,000 replicates are shown. (**b**) Statistical parsimony representation of the same 14 complete FMDV sequences, constructed using TCS (Version 1.21; ***34***). Each line represents a nucleotide substitution and each dot a putative ancestor virus.

depicting directly ancestral sequences (*see* **Fig. 3b**). Although the statistical parsimony trees drawn by TCS are not bootstrapped, if the data comprise the complete genome sequences of the sampled viruses, then the tree is as accurate and as representative as it can be: It is not sensitive to the choice of a single arbitrary locus because there are no further genetic data retrievable. A useful Web site that lists available phylogenetic programs for analyzing sequence data is http://evolution.genetics.washington.edu/phylip/software.html.

3.5. Future Technologies

Newer technologies are currently being developed that offer the potential to eliminate the use of capillary electrophoresis and even greater throughput. Resequencing microarrays have been developed and used to determine the sequence of the severe acute respiratory syndrome (SARS) coronavirus (***35,36***). However, development of specific arrays is heavily resource dependent and currently likely to be deployed only in niche markets. Of the newer technologies, sequential ligation systems (SOLiD), solid-phase primer amplification (Solexa), and bead-and-well-based pyrosequencing methods (such as the 454 platform) have the

capacity to generate reads of 4–20 Mb in a single run. Although this might be considered excessive for characterization of individual viral genomes, these approaches may allow infrequent mutations within a viral population to be detected. Thus, these methods may be ideal for dissecting the genetic variability within viral populations.

4. Notes

1. In addition to ensuring that all solutions used for RNA extraction are RNase free, pipets and work surfaces should be cleaned using 10% bleach followed by DNAzap (Ambion) prior to and between each sample processed.
2. A logical work flow for processing the samples for sequencing projects is highly recommended (*see* **Fig. 4** for an example). This is particularly important for high-resolution molecular epidemiological studies since the discrimination of samples may be dependent on the accurate determination of only a

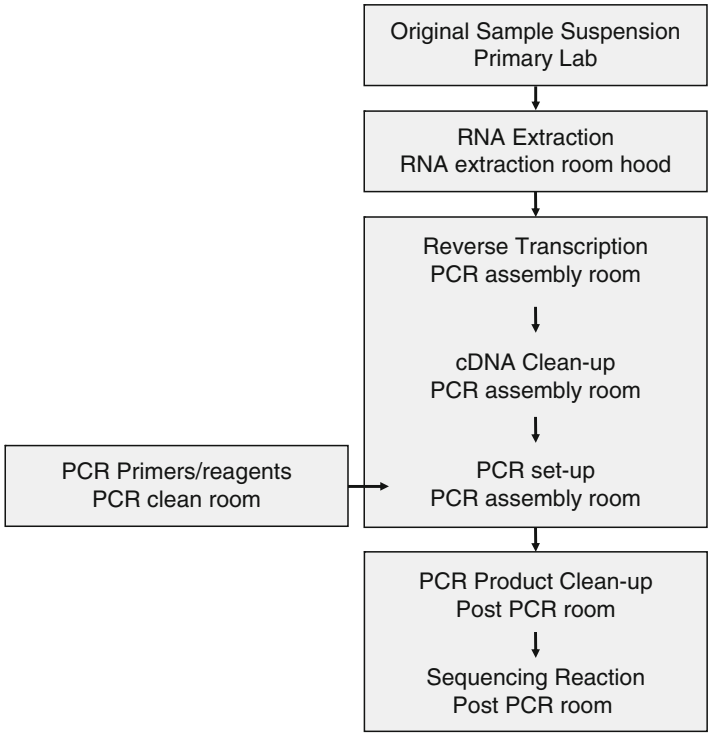


Fig. 4. Physical separation of the laboratory activities: outline of the separate steps required for the amplification and sequencing of FMDV genomes.

few nucleotide differences in the complete genome length *(3)*. Therefore, it is important that care is taken to minimize cross-contamination between samples (particularly post-PCR products). If possible, samples should be processed independently (including suitable negative control material), and the study should be organized to attempt to maximize the differences between successive samples tested.

3. A variety of sample types (including blood, tissues, esophageal-pharyngeal fluid, and cell culture supernatant) can be tested; however, it is usually preferable to test primary material (such as clinical samples) since it is possible that cell culture passage or molecular cloning of viruses can introduce nucleotide changes that can influence the interpretation of results.
4. Once placed in TRIzol reagent, samples can be stored for extended periods (at a wide range of temperatures, –70 to +4°C).
5. The requirement to perform a high number of downstream sequencing reactions may necessitate that a relatively large volume of PCR product is generated requiring pooling of RNA, cDNA, or post-PCR products. An additional practical consideration is the fidelity of the DNA polymerase used for the PCR amplification step; if possible proofreading enzymes that are widely available should be used.
6. In common with other long PCR methods, the parameters of the protocol used for amplification of viral genomes should be optimized prior to routine use. Steps to be considered include the components of the RT or PCR mixes and the cycling times used for amplification. In initial experiments, a PCR targeting a fragment of the 5′ end of the genome can be used to confirm that full-length cDNA has been produced in the RT reaction.
7. In general, these methods provide an accurate estimation of the viral consensus sequence. However, it is important to recognize that this sequence will be a composite of the component variability that, to a greater or lesser extent, may be present. In spite of concerns that it is theoretically possible that the sequence generated will not represent an actual virus species present in the sample, studies with FMDV indicated that the majority of molecular clones have identical sequences to the consensus *(37)*. Testing of duplicate samples can generate identical results *(4)*, demonstrating that these methods are accurate, and as long as the viral concentrations are relatively high, consensus sequences obtained will mask any individual proofreading errors that might arise due to low fidelity of reverse transcriptase and polymerase enzymes. These aspects relating to accurate determination of the sequences of specific viral genomes (rather than consensus sequences) will be of particular concern in studies

that aim to characterize the genetic population structure within samples (i.e., the quasi-species nature of a virus). New technologies and approaches (*see* **Subheading 3.5**.) may be utilized to address these important questions that underpin our understanding of viral evolution.

Acknowledgments

This work was funded by Defra research project SE2936. We acknowledge the assistance of colleagues Guido König, Sasmita Upadhyaya, Nigel Ferris, and Geoff Hutchings and Michael Quail from the Wellcome Trust Sanger Institute, Cambridge, for collaboration with the shotgun sequencing approach.

References

1. Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U S A* **74**, 5463–5467.
2. Carrillo, C., Tulman, E. R., Delhon, G., Lu, Z., Carreno, A., Vagnozzi, A., et al. (2005). Comparative genomics of foot-and-mouth disease virus. *J. Virol.* **79**, 6487–6504.
3. Cottam, E. M., Haydon, D. T., Paton, D. J., Gloster, J., Wilesmith, J. W., Ferris, N. P., et al. (2006). Molecular epidemiology of the foot-and-mouth disease virus outbreak in the United Kingdom in 2001. *J. Virol.* **80**, 11274–11282.
4. Cottam, E. M., Thébaud, G., Wadsworth, J., Gloster, J., Mansley, L., Paton, D. J., et al. (2008). Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc. Biol. Sci.* **275**, 887–895.
5. Carrillo, C., Lu, Z., Borca, M. V., Vagnozzi, A., Kutish, G. F., and Rock, D. L. (2007). Genetic and phenotypic variation of foot-and-mouth disease virus during serial passages in a natural host. *J. Virol.* **81**, 11341–11351.
6. Heath, L., van der Walt, E., Varsani, A., and Martin, D.P. (2006). Recombination patterns in aphthoviruses mirror those found in other picornaviruses. *J. Virol.* **80**, 11827–11832.
7. Jackson, A. L., O'Neill, H., Maree, F., Blignaut, B., Carrillo, C., Rodriguez, L., et al. (2007). Mosaic structure of foot-and-mouth disease virus genomes. *J. Gen. Virol.* **88**, 487–492.
8. Simmonds, P., Tuplin, A., and Evans, D. J. (2004). Detection of genome-scale ordered RNA structure (GORS). in genomes of positive-stranded RNA viruses: implications for virus evolution and host persistence. *RNA* **10**, 1337–1351.
9. Domingo, E., Escarmis, C., Lazaro, E., and Manrubia, S. C. (2005). Quasispecies dynamics and RNA extinction. *Virus Res.* 107, 129–139.
10. Cottam, E. M., Wadsworth, J., Shaw, A. E., Rowlands, R. J., Goatley, L., Maan, S., et al. (2008). Transmission pathways of foot-and-mouth disease virus in the United Kingdom in 2007. *PLoS Pathog.* **4**, e1000050.
11. Mason, P. W., Pacheco, J. M., Zhao, Q. Z., and Knowles, N. J. (2003). Comparisons of the complete genomes of Asian, African and European isolates of a recent foot-and-mouth disease virus type O pandemic strain (PanAsia). *J. Gen. Virol.* **84**, 1583–1593.
12. Carrillo, C., Tulman, E. R., Delhon, G., Lu, Z., Carreno, A., Vagnozzi, A., et al. (2006). High throughput sequencing and comparative genomics of foot-and-mouth disease virus. *Dev. Biol. (Basel)* **126**, 23–30.
13. Oem, J. K., Lee, K. N., Cho, I. S., Kye, S. J., Park, J. H., and Joo, Y. S. (2004). Comparison and analysis of the complete nucleotide sequence of foot-and-mouth disease viruses from animals in Korea and other PanAsia strains. *Virus Genes* **29**, 63–71.
14. Nobiron, I., Remond, M., Kaiser, C., Lebreton, F., Zientara, S., and Delmas, B. (2205). The nucleotide sequence of foot-and-mouth

disease virus O/FRA/1/2001 and comparison with its British parental strain O/UKG/35/2001. *Virus Res.* **108**, 225–229.

15. Du, J., Chang, H., Cong, G., Shao, J., Lin, T., Shang, Y., et al. (2007). Complete nucleotide sequence of a Chinese serotype Asia1 vaccine strain of foot-and-mouth disease virus. *Virus Genes* **35**, 635–642.
16. Li, X., Xu, Z., He, Y., Yao, Q., Zhang, K., Jin, M., et al. (2006). Genome comparison of a novel classical swine fever virus isolated in China in 2004 with other CSFV strains. *Virus Genes* **33**, 133–142.
17. Herring, B. L., Bernardin, F., Caglioti, S., Stramer, S., Tobler, L., Andrews, W., et al. (2007). Phylogenetic analysis of WNV in North American blood donors during the 2003–2004 epidemic seasons. *Virology* **363**, 220–228.
18. Marston, D. A., McElhinney, L. M., Johnson, N., Müller, T., Conzelmann, K. K., Tordo, N., et al. (2007). Comparative analysis of the full genome sequence of European bat lyssavirus type 1 and type 2 with other lyssaviruses and evidence for a conserved transcription termination and polyadenylation motif in the G-L 3′ non-translated region. *J. Gen. Virol.* **88**, 1302–1314.
19. Lu, L., Li, C., Fu, Y., Gao, F., Pybus, O. G., Abe, K., et al. (2007). Complete genomes of hepatitis C virus (HCV). subtypes 6c, 6l, 6o, 6p and 6q: completion of a full panel of genomes for HCV genotype 6. *J. Gen. Virol.* **88**, 1519–1525.
20. Bialasiewicz, S., Whiley, D. M., Lambert, S. B., Wang, D., Nissen, M. D., and Sloots, T. P. (2007). A newly reported human polyomavirus, KI virus, is present in the respiratory tract of Australian children. *J. Clin. Virol.* **40**, 15–18.
21. Jacobs, G. B., Loxton, A. G., Laten, A., and Engelbrecht, S. (2007). Complete genome sequencing of a non-syncytium-inducing HIV type 1 subtype D strain from Cape Town, South Africa. *AIDS Res. Hum. Retroviruses* **23**, 1575–1578.
22. Borges, M. B., Caride, E., Jabor, A. V., Malachias, J. M., Freire, M. S., Homma, A., et al. (2008). Study of the genetic stability of measles virus CAM-70 vaccine strain after serial passages in chicken embryo fibroblasts primary cultures. *Virus Genes* **36**, 35–44.
23. Xi, D., Li, J., Han, C., Li, D., Yu, J., and Zhou, X. (2008). Complete nucleotide sequence of a new strain of Tobacco necrosis virus A infecting soybean in China and infectivity of its full-length cDNA clone. *Virus Genes* **36**, 259–266.
24. Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376.
25. Huelsenbeck, J. P., Ronquist, F., Nielsen, R., and Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310–2314.
26. Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.
27. Michener, C. D., and Sokal, R. R. (1957). A quantitative approach to a problem in classification. *Evolution* **11**, 130–162.
28. Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174.
29. Lanave, C., Preparata, G., Saccone, C., and Serio, G. (1984). A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* **20**, 86–93.
30. Posada, D., and Crandall, K. A. (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817–818.
31. Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791.
32. Yang, Z., and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Mol. Biol. Evol.* **14**, 717–724.
33. Galtier, N., Gouy, M., and Gautier, C. (1996). SeaView and Phylo_win, two graphic tools for sequence alignment and molecular phylogeny. *Comput. Applic. Biosci.* **12**, 543–548.
34. Clement, M., Posada, D., and Crandall, K. A. (2000). TCS: a computer program to estimate gene genealogies. *Mol. Ecol.* **9**, 1657–1659.
35. Sulaiman, I. M., Liu, X., Frace, M., Sulaiman, N., Olsen-Rasmussen, M., Neuhaus, E., et al. (2006). Evaluation of affymetrix severe acute respiratory syndrome resequencing GeneChips in characterization of the genomes of two strains of coronavirus infecting humans. *Appl. Environ. Microbiol.* **72**, 207–211.
36. Wong, C. W., Albert, T. J., Vega, V. B., Norton, J. E., Cutler, D. J., Richmond, T. A., et al. (2004). Tracking the evolution of the SARS coronavirus using high-throughput, high-density resequencing arrays. *Genome Res.* **14**, 398–405.
37. Cottam, E. M., King, D. P., Wilson, A., Paton, D. J., and Haydon, D. T. (2009) Analysis of Foot-and-mouth disease virus nucleotide sequence variation within naturally infected epithelium. *Virus Res.* In press

Chapter 18

Bacterial Genome Sequencing

Hervé Tettelin and Tamara Feldblyum

Abstract

For over 30 yr, the Sanger method has been the standard for DNA sequencing. Instruments have been developed and improved over time to increase throughput, but they always relied on the same technology. Today, we are facing a revolution in DNA sequencing with many drastically different platforms that have become or will soon become available on the market. We review a number of sequencing technologies and provide examples of applications. We also discuss the impact genomics and new DNA sequencing approaches have had on various fields of biological research.

Key words: Bacteria, diversity, genome, pathogen, sequencing technology.

1. Introduction

1.1. The Advent of Whole-Genome Shotgun Sequencing

Until the completion of the genome sequence of *Haemophilus influenzae* Rd *(1)* by the whole-genome shotgun approach, genome sequencing projects relied on the availability of a physical map of the organism of interest *(2)*. Such a map is a set of partially overlapping clones from a genomic library, usually with large inserts like cosmids (ca. 40 kb) or bacterial artificial chromosomes (BACs) (ca. 100 kb), which are ordered along the genome. The construction of physical maps typically involves fingerprinting of a large number of clones and assembly of the map based on fingerprint overlaps. Once the map is assembled, a subset of individual clones tiling the genome is selected and used for sequencing. Although physical maps or derivatives thereof are still currently used to increase the accuracy of eukaryotic genome sequencing projects that involve complex repeats, most of the whole-genome

D.A. Caugant (ed.), *Molecular Epidemiology of Microorganisms*, Methods in Molecular Biology, Vol. 551
DOI: 10.1007/978-1-60327-999-4_18,

sequencing now is conducted using the shotgun approach. This technique involves the sequencing of a large number of clones chosen randomly from a whole-genome library. Such clones typically have small-to-medium-size inserts (2–10 kb) and are often sequenced from both ends of the insert (paired ends) to provide scaffolding of repeats. Random sequences are then aligned and assembled together into contigs to reconstruct the structure of the genome. Resulting assemblies typically consist of a number of contigs separated by gaps in which sequence information is missing, most often due to complex repeats that are hard to assemble, DNA regions that are difficult to sequence (e.g., secondary structures), or fragments missing from the genomic library (unclonable regions). Gap closure is usually laborious because it inevitably tackles the most difficult regions of the genome to sequence, and it involves a variety of molecular biology techniques, depending on the case at hand *(3)*.

Whether a whole bacterial genome sequence obtained is complete (gap free) or a draft (a set of contigs separated by gaps), the genes it harbors can be systematically predicted using tools such as Glimmer *(4)* or GeneMark *(5)*. The function of the proteins encoded by those genes can be predicted in 60 to 90% of the cases by homology to characterized proteins available in the ever-growing public databases. Predicted coding regions are typically searched against public databases with BlastP *(6)* and assigned an annotation as well as a functional role category *(7)* or Gene Ontology (GO) terms *(8)*. To further enhance function assignment, the proteins are also searched against databases of hidden Markov models (HMMs) built on protein family/superfamily multiple-sequence alignments *(9,10)*. In addition to *ab* initio prediction of coding regions, comparative genomics is used to drive and improve annotation as well as to make it more homogeneous across strains or species. These steps are just some of the basic aspects of a full-blown genome annotation pipeline whose description is beyond the scope of this chapter.

The genomics field is now increasingly turning towards metagenomics that applies genomics techniques to the study of complex communities of microbial organisms directly in their natural environments and without the need of laboratory isolation and cultivation. It aims to capture the total microbial gene diversity in an environment, shedding light on the biological processes found there. Examples of metagenomics projects include environmental studies as those of seawater *(11)* or soil *(12)*, and medical applications, such as the human microbiome project *(13)*.

1.2. Completed and Ongoing Bacterial Genome Projects

As of February 2009, the Genomes Online Database (GOLD; www.genomesonline.org) reports 792 complete published bacterial genomes and another 2,392 ongoing bacterial genome

projects. These include all the major human pathogens, a growing number of other pathogens, and bacteria of environmental and industrial relevance. Such a flood of genomics data requires the design and access to databases that enable interrogation of the information in a biological context. Some databases like the Comprehensive Microbial Resource (cmr.jcvi.org) aim at providing comparative power across a comprehensive list of completely sequenced species. Other databases target a subset of species like the Bioinformatics Resource Centers (www.brc-central.org). The Bioinformatics Links Directory (bioinformatics.ca/links_directory) features a long list of links to molecular resources, tools, and databases *(14)*. This directory provides an excellent starting point for users to get acquainted with the most useful and powerful publicly available tools for genomic data mining and analysis.

2. Sequencing Technologies

2.1. Existing Technologies

Ever since whole-genome shotgun sequencing became the standard, the intrinsic sequencing approach did not change until recently. Although several generations of improved automated sequencers were developed and the speed of shotgun sequencing increased 25 times between 1986, when the first automated DNA sequencer (Applied Biosystems model 370A) was commercialized, and 2005, when the AB 3730xl capillary sequencer dominated the laboratories, the steps in the high-throughput shotgun sequencing process did not evolve. Genomic DNA was fragmented into pieces of 2–40 kb and genomic libraries constructed by cloning the fragments into plasmid or fosmid vectors and transforming the constructs into *Escherichia coli* for replication and propagation. The plasmid or fosmid DNA was then isolated and used as the sequencing template for dideoxy-mediated chain termination sequencing reactions. The dideoxy-mediated chain termination sequencing chemistry has been the standard in the field since its discovery and publication in 1975 by Fred Sanger *(15,16)*.

The way DNA sequence information is generated was revolutionized in 2005 by 454 Life Sciences with the release of a sequencing platform, the Genome Sequencer 20 (www.roche-applied-science.com) that is based on totally different chemistry and technology *(17)* (*see* **Table 1** for a summary of the sequencing technologies described in this section). It does not require cloning of the shotgun fragments and therefore eliminates the cloning bias for fragments that were unstable or could not be propagated in *E. coli*.

The 454 sequencing sample preparation steps include DNA fragmentation, end repair, capture of the fragments on beads,

Table 1
Characteristics of Sequencing Technologies

Sequencer	Sequencing chemistry	Sample preparation	Estimated throughput	Read length (basepairs)	Accuracy base read (consensus)
ABI 3730xl DNA Analyzer	Dideoxy termination	8 d	0.08 Mb/run 1 Mb/d	800–1,000	>99%
454-Roche Genome Sequencer FLX (Titanium)	Sequencing by synthesis, Pyrosequencing	5 d	400–600 Mb/run 10 h	400	>99.5% (99.99% at ~20X coverage)
Illumina-Solexa 1G Genome Analyzer$_{II}$ System	Sequencing by synthesis, Reversible terminators	1 d	2–15 Gb/run 2–8 days	35–75	98–99 (99.99% at > 3X coverage)
ABI SOLiD 2	Stepwise ligation	2–4 d	16 Gb/run 6–8 days	35	99.94% (99.999% at 15X coverage)
Helicos BioSciences Heli-Scope Single Molecule Sequencer	Sequencing by synthesis	1 day	21–28 Gb/run 8 days	30–35	>95% [99.995% at >20X coverage]

Note: This table provides information on the technology, throughput, read length, and accuracy of platforms on the market as of December 2008 (described in **Subheading 2.1.**).

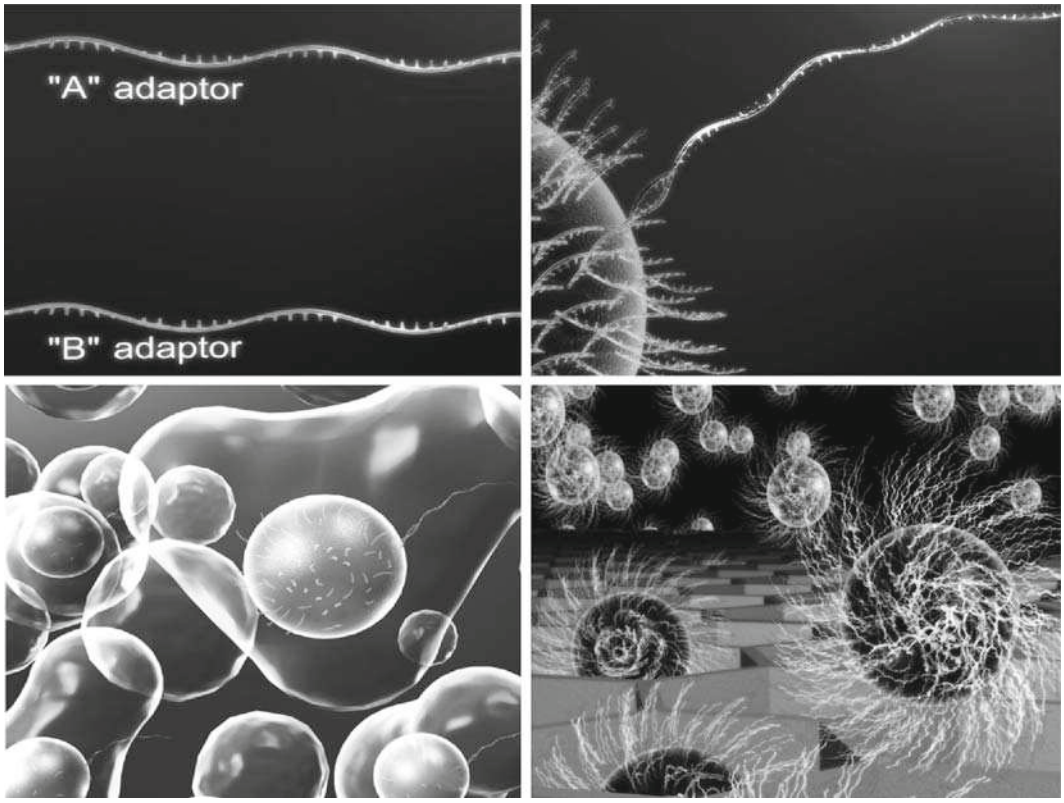


Fig. 1. The 454 sequencing steps: ligation of adapters, capture of DNA fragments on beads, and clonal amplification by PCR in emulsion microreactors.

polymerase chain reaction (PCR) clonal amplification of the captured fragments in aqueous-oil emulsion microreactors, breaking of the microreactors, and enrichment for beads with amplified DNA (**Fig. 1**). The beads are then loaded into wells on a picotiter plate that is placed on the surface of a charge-coupled device (CCD). The sequencing is performed by synthesis using a modified pyrosequencing *(18)* procedure on solid support (**Fig. 2**). Nucleotides are sequentially passed through the flow chamber, and complementary nucleotides are incorporated in the wells containing the template-carrying beads. Inorganic pyrophosphate and photons are generated during the synthesis, and the signal from the individual wells is captured by the CCD, enabling reading of the template sequence.

The initial read length for each template was about 100 bases, which is much shorter than the average 800 bases read length routinely achieved by Sanger sequencing. It was improved to an average of 250 bases with the introduction of a second-generation 454 FLX sequencer and has reached over 400 bases for a total of 1–2 Gb in 24 hours with further improved consumables and software upgrades (Titanium series). Paired-end

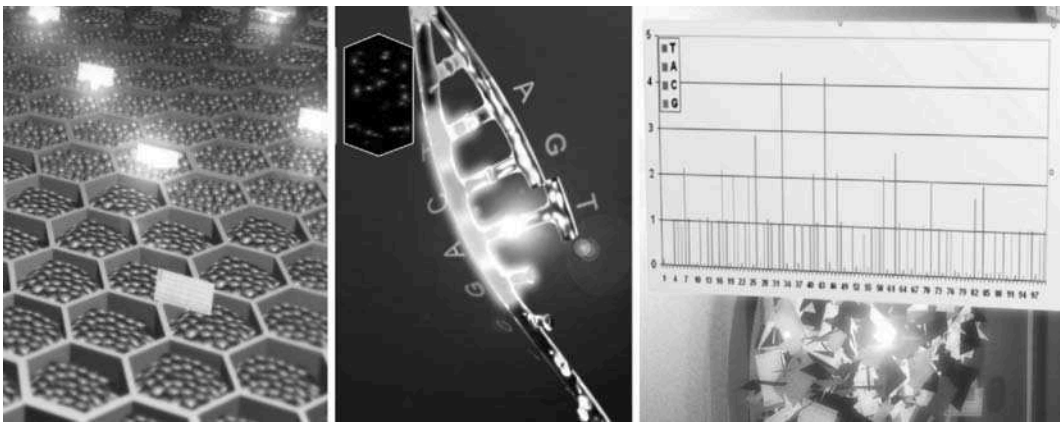


Fig. 2. The 454 sequencing steps: loading of picotiter plate, pyrosequencing, and reading of the flowgram.

sequencing protocol that generates 110 base-pair tags separated by 3000 bp genomic distance and that mimics the Sanger double-strand sequencing was released in 2008. This approach will improve the *de novo* sequence assembly and may overcome one of the drawbacks of the technology, namely, the inability to correctly assemble large repeats. The other main drawback relates to homopolymeric tracts, especially in monomers longer than three or four bases, for which the quantitation of the signal is not always accurate because all the bases in the tract are added in one nucleotide flow (e.g., six A's in a row) and generate a single signal of higher intensity. This problem is partially overcome by the generation of a much deeper sequence coverage of each base than is traditionally achieved with Sanger sequencing. Each base is sequenced approx 30 times, instead of only 5–8 times, in a single 454 run for an average bacterial genome.

Another popular sequencing instrument is the Illumina Genome Analyzer (www.illumina.com), released in June 2006. It employs the Solexa technology that is based on sequencing by synthesis, reversible terminator chemistry, and clonal single-molecule array technology. A DNA sample of 100 ng/μg is fragmented, and approximately 40 million random DNA fragments are immobilized on the optically transparent planar surface of a flow cell at a density of up to 10 million/cm^2. The fragments are amplified (solid-phase amplification) to create close to 1,000 local copies of each fragment. All four fluorescently labeled modified nucleotides possessing the reverse termination property and a polymerase are added to the array for sequencing. A laser light of a specific wavelength for each base excites the label on the incorporated nucleotides, and fluorescence is detected by a CCD. The fluorophore and the reversible terminator are then removed, and the cycle of incorporation, detection, and identification is repeated. A sequencing run generates up to 1 Gb

in random short sequence reads of 30–50 bases, producing an evenly distributed high coverage of reads that can be aligned with a reference genome sequence. The scalable nature of this platform allows for a paired-end run on a single flow cell generating up to 2 billion bases in a 6-day run. The presence of all four reversible-terminator deoxynucleotide 5′-triphosphates (dNTPs) during each sequencing cycle minimizes the incorporation bias and increases the accuracy of base calls. The quality of each base call is determined independently; therefore, sequencing through homopolymeric regions poses no additional challenge.

The ABI SOLiD system, which is based on sequencing by sequential ligation, was released by Applied Biosystems in October 2007 (www.appliedbiosystems.com). The technology combines massively parallel clonally amplified sample preparation, sequencing by ligation, and fluorescently tagged nucleotide detection. Up to 10 μg of starting material is fragmented, the 60–90 base DNA pieces are linked to magnetic beads and clonally amplified in emulsion PCR microreactors (similar to the 454 protocol). The beads containing amplified DNA are covalently bound to glass slides and loaded onto the SOLiD Analyzer. The sequence of the fragments is generated by ligation with four-color dye-labeled probes. Every fourth and fifth base is interrogated; the probe with exact matching bases ligates to the bound fragment, the color signal is detected and recorded at the fifth base, and the fluorescent tag is cleaved. After seven cycles of ligation, the original primer is stripped from the template, and a new primer offset by one base is hybridized to begin the next series of ligations. Sequences of the first 25–35 bases are determined after five rounds of primer resets. The technology is also developed for paired-end sequencing across a range of library insert sizes from 0.6 to 10 kb. Interrogating two adjacent bases simultaneously during the sequencing process improves the base call accuracy and may allow distinction between sequencing errors and single-nucleotide polymorphisms (SNPs). Currently the SOLiD Analyzer generates about 2 Gb of data per run for a fragmented library and about 4 Gb for a paired-end library sequenced on two slides.

The HeliScope instrument from Helicos Biosciences (www.helicosbio.com) is the first commercialized single-molecule sequencing platform that does not require DNA amplification. The technology is based on the work of Steven Quake ***(19)***, who developed a novel surface chemistry that anchored DNA molecules to microchannels and demonstrated the proof of principle that the sequence of a template can be revealed by sequential addition of one type of labeled nucleotide and DNA polymerase. The Helicos sequencing by synthesis is performed in two flow cells in which single-stranded DNA molecules are captured on a

chemically treated surface at a density of 100 million/cm^2. The DNA polymerase and fluorescently labeled bases are sequentially added, and if the base is complementary to the template, it is incorporated into the synthesized strand. Unincorporated bases are removed, the positions of the fluorescent bases are captured, and the fluorescent tag is cleaved. While a base is added in one cell the image is captured in the other cell. The first-generation HeliScope systems released in early 2008 produce 25–55 base reads with <5% raw error rate for a total capacity of the instrument estimated to be about 1–3 Gb/d. The instrument has the potential to generate up to 10^9 bases per hour with future improvements in chemistries and the flow cell. Single-molecule sequencing without DNA template amplification eliminates the risk of introducing errors during the amplification step but increases the probability of a wrong base call with a single-molecule template because of base misincorporations by the polymerase. This technology has the potential to considerably reduce the sequencing costs because of its massively parallel nature and minimal use of reagents if the data generated prove to be highly accurate.

2.2. Combination of Sequencing Technologies

The combination of different technologies produces the most complete genome sequences. Disadvantages of one platform are complemented by the ability to resolve these features with a different technology. For example, sequences that cannot be propagated in *E. coli* are not represented in the cloned libraries, resulting in sequence gaps in shotgun assemblies. The steps of cloning and propagation in the bacterial host are absent in the 454 sample preparation work flow. The fragmented genomic DNA is amplified on beads in emulsion PCR; therefore no host biases affect the composition of the amplified library. On the other hand, the pyrosequencing chemistry utilized on the 454 platform exhibits limited capability in accurately resolving homopolymers longer than three or four bases, whereas the Sanger dideoxy-terminator sequencing produces the accurate number of consecutively repeated bases. The hybrid assembly of shotgun reads generated by these two platforms results in an accurate and almost complete microbial genome sequence, reducing, and in some cases eliminating, the need for manual genome finishing ***(20)***. The 454 platform is also an effective tool for sequencing through areas of secondary structures (hard stops) that are often prevalent in genomes with high G + C content (>60% G + C).

The important factors are not only the total bases produced by a sequencer, but also the read length obtained from each sequenced fragment. It is especially significant in the case of *de novo* sequencing and assembly of the shotgun reads. Addition of a large number of relatively short pyrosequencing reads did not result in significant reduction in gaps in genomes containing

a large number of repetitive areas or in the number of physical (unlinked) gaps *(20)*. The use of paired ends is therefore necessary in this kind of genome.

McCutcheon and Moran *(21)* demonstrated that a complete and accurate genome sequence of *Sulcia muelleri* can be obtained by combining two of the new higher-throughput technologies. They mapped short accurate reads (33 bases) from a partial Solexa Genome Analyzer run onto assembled 454 data. In doing so, they successfully corrected the 454 sequencing errors in the homopolymeric regions and eliminated the majority of the frameshifts in coding regions.

2.3. Future Technologies

In addition to the sequencing platforms released in the last 2–3 years, numerous other companies and academic groups continue their quest for a technology that will yield a $1,000 human genome *(22)* and bring sequencing closer to clinical applications.

Intelligent Bio-Systems (www.intelligentbiosystems.com) is developing a Pinpoint sequencer with predicted capacity of several gigabasepairs per day. It is based on sequencing by synthesis of PCR-amplified DNA fragments captured on a high-density glass chip. The modified nucleotides, containing an end cap and labeled with a base-specific removable fluorescent dye, are incorporated during the DNA strand synthesis. The array is scanned; the fluorescent label on the terminal base is detected, recorded, and then cleaved along with the end cap.

The VisiGen Biotechnologies sequencing chemistry (www.visigenbio.com) is based on real-time single-molecule fluorescence detection in a massive parallel array. The DNA polymerase is modified with a fluorescent donor, and each nucleotide is color coded with an acceptor fluorescent tag. During the extension reaction, when a nucleotide is incorporated into the growing DNA strand, energy transfers from the polymerase to the nucleotide (fluorescence resonance energy transfer, FRET), and a base-specific signal is emitted and detected in real time. The goal of VisiGen is to achieve a sequencing rate of 1 Mb/s per machine and generate read lengths of about 1,000 bases.

Mobious (www.mobious.com) is combining biological molecular systems and artificial nanostructures to create novel single-molecule sequencing and array-based sequencing-by-synthesis platforms. Mobious's approach circumvents the use of labeled nucleotides and secondary enzymes. Known as polykinetic sequencing, it takes advantage of the selective mechanisms occurring during the polymerase reaction, discriminating between the time the DNA polymerase takes to add a complementary base to a growing strand and the time it takes to reject a noncomplementary base. By labeling the polymerase with a fluorescent tag

and adding a single nucleotide at a time, the sequence is derived by measuring the amount of time the polymerase attaches to the growing strand.

NABsys (www.nabsys.com) in partnership with Brown University is developing the hybridization-assisted nanopore sequencing (HANS) platform. The genomic DNA is fragmented in 100-kb pieces and made single stranded. Oligonucleotides representing all possible permutations of six bases are attached to the fragments one at a time. The fragments with annealed oligonucleotides are then passed through nanopores, and the current flow through the pore is recorded. A drop in current indicates the presence of an oligonucleotide. The profiles of 6-mer positions along each fragment are then assembled into the 6-mer map of the entire genome. Finally, the profiles of each possible 6-mer are aligned, and the whole-genome sequence is derived.

Pacific Biosciences (www.pacificbiosciences.com) is exploring a different approach to sequencing a single native DNA molecule in real time. The SMRT™ sequencer utilizes a zero-mode waveguide based on Harold Craighead's and Stephen Turner's work at Cornell University *(23,24)*. The zero-mode waveguides are nanometer-scale holes in a very thin metal film in which a single DNA polymerase molecule is captured. The sequence is determined during the DNA replication in real time as the DNA polymerase adds nucleotides, each tagged with a different fluorescent dye, to the growing complementary DNA (cDNA) strand. Because light can penetrate only a very short distance past the hole, the imaging equipment illuminates just the base added. The sensor can detect only that spotlight, but not all other free-floating fluorescent bases. The current reported throughput of the SMRT system is about 10 kb/s or 36 Mb/h, but projected improvements could bring the sequencing speed to 100 Gb/h. If the promise of this technology is realized, the human genome could be sequenced in about 4 min for less than $1,000.

The goal of Reveo (www.reveo.com/vision), developers of the OmniMoRA (omni molecular recognizer application), is to sequence a human genome in less than 1 min. They rely on physical electrooptic methods and nanotechnology rather than the traditional chemical methods. Sequencing is performed by scanning stretched immobilized single-strand DNA using an array of nano-knife-edge blades as detectors. Accelerated electrons excite the bases, which vibrate with specific frequencies, and the molecular vibration characteristics are measured and recorded for each of the nucleotides. The same device could sequence amino acids in a protein. It has the potential to achieve improvements over existing sequencing instruments that may lead to 100% error-free reads and 100% coverage of the human genome in minutes for pennies per genome.

3. Applications

Until future technologies become readily available, the increase in sequencing throughput will come with shorter read lengths or other varying drawbacks. Although the combination of different technologies in a single project can alleviate some of the drawbacks and improve sequencing results (*see* **Subheading 2.2**), newer technologies can be used individually in applications that go beyond genome *de novo* sequencing.

The most obvious use of a very large number of short reads is to align them onto a reference complete genome to identify differences, in particular small local ones like SNPs. The ability to reliably distinguish valid SNPs from sequencing errors depends on the average quality of the reads generated by the technology as well as its throughput and ability to perform several runs easily, which directly correlates to the average number of individual reads that will span any given SNP.

If paired ends are combined with very high-throughput sequencing technologies, the rapid mapping of an extremely large number of sequence pairs onto a reference genome enables the identification and characterization of rearrangements, such as deletions, insertions, and inversions, that happened between the pairs. For instance, the 454 technology was used to detect and map more than 1,000 structural variations, 3 kb or larger, between two humans using the human genome as a reference *(25)*.

The higher throughput of new sequencing platforms also positions them favorably for the detection of rare entities or events. For instance, 454 sequencing was used to generate expressed sequence tags (ESTs) from RNA captured by laser microdissection from the shoot apical meristem of maize. In a single run, over 25,000 genes were tagged, and 30% of the 454 ESTs did not align with ESTs previously generated for this tissue. The 454 ESTs included rare transcripts that had not yet been captured *(26)*. Sequencing cDNA from RNA samples using these technologies also provides for accurate determination of the relative abundance of individual RNAs within the sample simply by comparing the relative abundance of sequences obtained for each gene/transcript. A study indicated that this approach combines the high-throughput advantage of serial analysis of gene expression (SAGE) with the mapping accuracy of EST sequencing that provides longer reads than the short SAGE tags *(27)*. The study also determined that short and long transcripts (<80 bases or >300–400 bases, respectively) are underrepresented in 454 sequencing reads. However, the limitation with longer transcripts can easily be overcome by shearing the starting material, for instance, by nebulization. Nevertheless, in another study 454

was used in conjunction with Solexa to detect small regulatory RNAs (20–30 bases), including microRNAs *(28)*. It is foreseeable that new sequencing technologies will make microarray-based comparative genomic hybridizations (CGHs) and transcriptional profiling obsolete in the near future given their higher accuracy than microarray hybridization results and broader dynamic range of detection. The Solexa technology has also been used for ChIP-Seq *(29)* to study protein-DNA interactions, for instance, to characterize promoter-binding sites; the identification of DNA methylation patterns in Arabidospsis *(30)*; and the mapping of nucleosome positions in humans *(31)*. The list of applications is as diverse as it is long, and it will continue to grow as additional technologies come online.

It should be noted that the use of these new technologies comes with significant challenges that are sometimes overlooked by interested users. First, the sheer throughput of the new platforms generates amounts of data that can rapidly become extremely difficult to handle, from the very basic aspect of storage on digital media to the need for robust software applications to process and analyze the data. Second, the disparity in sequencing accuracy, read length, and read type (e.g., Sanger electropherograms vs. 454 flowgrams) renders the assembly of reads from individual platforms or combinations thereof quite challenging. Tools are currently being developed to address this issue *(32)*, and it is hoped that they will enable users to select technologies that best suit their needs and generate reliable assemblies for further analysis.

4. Impact of Whole-Genome Sequencing

The availability of genome sequence information and tools to mine it has revolutionized the way researchers in many fields design their experiments. Molecular biology is probably the most affected discipline as manipulation of genes has become so powerful with the knowledge of their sequence. Despite the large number of species sequenced, genomics continues to unravel genes that had not been seen before and whose function is unknown. One of the goals of the genomic art is to point experts toward a set of new genes that are of interest to their research. For instance, the context of these genes—operons, distribution across species or strains thereof (comparative genomics), or phylogenetic trees—can help gain insights into their potential role and open avenues for research. Subsequently, the identification of novel gene functions will lead to new research applications.

Genomics has shed light on many aspects of evolution *(33)*: genome reduction as seen in the case of obligate intracellular bacteria; genome plasticity (rearrangements, mobile elements); gene duplication and diversification of protein function; lateral gene transfer and acquisition of new functions; adaptation to environments; virulence; and so on. It has also had an impact on industrial processes, bioremediation, and biotransformation, as well as medicine with the accelerated development of vaccines (*see* **Subheading 5**.), drugs, and diagnostics. Epidemiology is of course intimately connected to genomics. The latter provides a whole-genome perspective to the classifications derived from the subsets of markers measured by molecular epidemiology techniques. On the other hand, epidemiology is excellent at identifying strains that should be selected for whole-genome sequencing to encompass and thoroughly characterize the breadth of diversity at hand.

Many new disciplines have also emerged or significantly expanded in the postgenomic era, including

1. Functional genomics that tackles the function of genes at the whole-genome level. Transcriptomics, using microarrays or cDNA sequencing techniques, identifies the transcriptional level of the entire gene repertoire under various conditions. For example, Grifantini et al. identified the genes expressed on interaction of *Neisseria meningitidis* with epithelial cells *(34)*. Proteomics achieves a similar goal but by looking at the protein level rather than the messenger RNA (mRNA) level. For instance, Pieper et al. studied the resistance of *Staphylococcus aureus* to vancomycin using comparative proteomics approaches *(35)*. Metabolomics tackles yet another level of the cell biology: the profile of all metabolites, the small molecules that are the end product of specific cellular processes *(36)*. Interactomics investigates the protein-protein interactions within the bacterial cell or between the bacterial proteins and their host, for instance, using two-hybrid techniques *(37)*. The difficult task of reconciling the knowledge gained from all these approaches belongs to the rising field of systems biology, in which studies are conducted at the level of whole cells or communities *(38)*.
2. Synthetic biology: Given the sequence of an entire genome, it is possible to synthesize genes *de novo*, usually starting from long synthetic oligonucleotides that are assembled together sequentially. A common application is the optimization of codon usage within a gene of interest for more efficient heterologous expression, for instance, for the production of a specific compound. More recently, investigators have attempted to identify the minimal genome, the smallest set of genes that enables life, and synthesized the minimal genome of *Mycoplasma genitalium (39,40)*.

3. Structural genomics: Obtaining the three-dimensional structure of proteins is the ultimate step toward characterizing their function and understanding their interaction with their environment. Comparative genomics facilitates the comprehensive identification of protein families representative of particular protein functions. Ongoing projects aim at systematically crystallizing a representative member of each of those families and deciphering their three-dimensional structure ***(41)***. Given the high degree of diversity among some protein families, subsequent efforts will likely be aimed at crystallizing multiple members of the more diverse families to shed light into the evolution of function.

5. Reverse Vaccinology and Bacterial Diversity

One of the main goals of sequencing the genome of many strains and species of bacterial pathogens is to identify novel tools to help combat disease. Reverse vaccinology ***(42)*** makes use of genomic sequence information to identify novel and better-suited protein candidates for vaccine development. Knowing the genome provides access to all the proteins it encodes and an understanding of their diversity, thus enabling a more informed selection of vaccine candidates.

Reverse vaccinology was pioneered in 2000 on serogroup B *Neisseria meningitidis* ***(43,44)***. Based on the genomic data, all proteins predicted to be surface exposed and therefore likely to be accessible to antibodies were identified *in silico*. Criteria for prediction included proteins known to carry out functions at the surface of the cell; exclusion of proteins known to be cytoplasmic; exclusion of proteins likely to be embedded in the cell's membrane and inaccessible to antibodies; and amino acid motifs characteristic of targeting to the membrane (signal peptides), anchoring in the lipid bilayer (lipoproteins), anchoring in the outer membrane of gram-negative bacteria or the cell wall of gram-positive bacteria, and interaction with host proteins or structures (e.g., integrin-binding domain) ***(45)***. This resulted in a list of approx 600 genes that were systematically cloned in *E. coli* for expression of recombinant proteins. About 350 proteins that were successfully expressed and purified were used for characterization of their antigenicity and accessibility on the cell surface. Of these, 85 were positive in one or more of the following assays: Western blot (specificity), flow cytometry or immunoprecipitation (accessibility), and bactericidal activity (ability of the antisera to kill the bacteria in vitro when combined with human complement) ***(46)***. The last is a good indicator that the antigen is a promising vaccine candidate. The seven best candidates that satisfied all criteria were selected

and sequenced across a panel of diverse strains of *N. meningitidis* representing all serotypes and spanning the phylogeny of the species ***(44)***. Five of the seven candidates were completely conserved across the entire panel of strains. Thus, for the first time in decades of classical vaccinology, five extremely strong vaccine candidates likely to confer general protection against serogroup B strains of *N. meningitidis* were identified. These were combined and tested in infant rats challenged intraperitoneally with lethal doses of *N. meningitidis*. The cocktail, when formulated with adjuvants suitable for human use, conferred protection in rats against 90% of a panel of 85 *N. meningitidis* strains representative of the global population diversity ***(47)***. The cocktail is currently being tested in human clinical trials ***(47)***.

Since then, the reverse vaccinology approach has been applied to numerous microbial species. Another striking example of its importance is the case of group B *Streptococcus (Streptococcus agalactiae)*. The first genome sequence of this species did not provide antigens able to confer general protection against the diversity of strains encountered in the clinic. The generation of eight complete genome sequences encompassing the major disease-causing serotypes indicated that *S. agalactiae* is a very diverse species. Indeed, each new genome sequence provided an average of 33 new genes, and mathematical extrapolation of the trend indicated that a very large number of genomes would have to be sequenced before the entire gene repertoire of group *B Streptococcus* could be determined ***(48,49)***. This led to the concept of the bacterial pan-genome, composed of the core genome: the genes present in all sequenced strains and the dispensable genome made of genes present in a subset of the strains. The latter contributes to the diversity of the species and provides the tools for evolution and adaptation. In the case of *S. agalactiae*, the pan-genome is described as open, meaning that the size of the pan-genome is undetermined and is likely to be very large. Other species, such as *Bacillus anthracis*, exhibit a closed pan-genome because only four genome sequences are sufficient to describe their entire gene repertoire ***(49)***. The pan-genome concept has deep implications for the diversity of the species and the discovery of vaccine candidates. In the case of *S. agalactiae*, a cocktail of four antigens, only one of which was part of the core genome, had to be used to confer broad protection ***(50)***.

Reverse vaccinology is only one example of the power of bacterial genome sequencing in the modern era of genomics. In the near future, a remarkably large number of bacterial species will have one or several genome sequences available (complete or draft), including unculturable species, thanks to metagenomics and other approaches. This wealth of data will continue to alter the way we conduct research and warrants an exciting future in our respective fields.

References

1. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.
2. Fonstein, M., and Haselkorn, R. (1995). Physical mapping of bacterial genomes. *J. Bacteriol.* **177**, 3361–3369.
3. Frangeul, L., Nelson, K. E., Buchrieser, C., Danchin, A., Glaser, P., and Kunst, F. (1999). Cloning and assembly strategies in microbial genome projects. *Microbiology* **145**, 2625–2634.
4. Delcher, A. L., Bratke, K. A., Powers, E. C., and Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673–679.
5. Besemer, J., and Borodovsky, M. (2005). GeneMark: Web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.* **33**, 451–454.
6. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
7. Riley, M. (1993). Functions of the gene products of *Escherichia coli*. *Microbiol. Rev.* **57**, 862–952.
8. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *The Gene Ontology Consortium. Nat. Genet.* **25**, 25–29.
9. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L., and Sonnhammer, E. L. (2000). The Pfam protein families database. *Nucleic Acids Res.* **28**, 263–266.
10. Haft, D. H., Selengut, J. D., and White, O. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373.
11. Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74.
12. Daniel, R. (2005). The metagenomics of soil. *Nat. Rev. Microbiol.* **3**, 470–478.
13. Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature* **449**, 804–810.
14. Fox, J. A., McMillan, S., and Ouellette, B. F. (2007). Conducting research on the web: 2007 update for the bioinformatics links directory. *Nucleic Acids Res.* **35**, 3–5.
15. Sanger, F., and Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94**, 441–448.
16. Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U S A* **74**, 5463–5467.
17. Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380.
18. Ronaghi, M., Uhlen, M., and Nyren, P. (1998). A sequencing method based on real-time pyrophosphate. *Science* **281**, 363–365.
19. Kartalov, E. P., and Quake, S. R. (2004). Microfluidic device reads up to four consecutive base pairs in DNA sequencing-by-synthesis. *Nucleic Acids Res.* **32**, 2873–2879.
20. Goldberg, S. M., Johnson, J., Busam, D., Feldblyum, T., Ferriera, S., Friedman, R., et al. (2006). A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc. Natl. Acad. Sci. U S A* **103**, 11240–11245.
21. McCutcheon, J. P., and Moran, N. A. (2007). Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. *Proc. Natl. Acad. Sci. U S A* **104**, 19392–19397.
22. Bennett, S. T., Barnes, C., Cox, A., Davies, L., and Brown, C. (2005). Toward the 1,000 dollars human genome. *Pharmacogenomics* **6**, 373–382.
23. Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G., and Webb, W. W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**, 682–686.
24. Korlach, J., Marks, P. J., Cicero, R. L., Gray, J. J., Murphy, D. L., Roitman, D. B., et al. (2008). Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc. Natl. Acad. Sci. U S A* **105**, 1176–1181.
25. Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., et al. (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426.
26. Emrich, S. J., Barbazuk, W. B., Li, L., and Schnable, P. S. (2007). Gene discovery and

annotation using LCM-454 transcriptome sequencing. *Genome Res.* **17**, 69–73.

27. Torres, T. T., Metta, M., Ottenwalder, B., and Schlotterer, C. (2008). Gene expression profiling by massively parallel sequencing. *Genome Res.* **18**, 172–177.
28. Hafner, M., Landgraf, P., Ludwig, J., Rice, A., Ojo, T., Lin, C., et al. (2008). Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods* **44**, 3–12.
29. Mardis, E. R. (2007). ChIP-seq: welcome to the new frontier. *Nat. Methods* 4, 613–614.
30. Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., et al. (2008). Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**, 215–219.
31. Schones, D. E., Cui, K., Cuddapah, S., Roh, T. Y., Barski, A., Wang, Z., et al. (2008). Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887–898.
32. Pop, M., and Salzberg, S. L. (2008). Bioinformatics challenges of new sequencing technology. *Trends Genet.* **24**, 142–149.
33. Fraser-Liggett, C. M. (2005). Insights on biology and evolution from microbial genome sequencing. *Genome Res.* **15**, 1603–1610.
34. Grifantini, R., Bartolini, E., Muzzi, A., Draghi, M., Frigimelica, E., Berger, J., et al. (2002). Previously unrecognized vaccine candidates against group B meningococcus identified by DNA microarrays. *Nat. Biotechnol.* **20**, 914–921.
35. Pieper, R., Gatlin-Bunai, C. L., Mongodin, E. F., Parmar, P. P., Huang, S. T., Clark, D. J., et al. (2006). Comparative proteomic analysis of *Staphylococcus aureus* strains with differences in resistance to the cell wall-targeting antibiotic vancomycin. *Proteomics* **6**, 4246–4258.
36. Oldiges, M., Lutz, S., Pflug, S., Schroer, K., Stein, N., and Wiendahl, C. (2007). Metabolomics: current state and evolving methodologies and tools. *Appl. Microbiol. Biotechnol.* **76**, 495–511.
37. Collura, V., and Boissy, G. (2007). From protein-protein complexes to interactomics. *Subcell. Biochem.* **43**, 135–183.
38. Kitano, H. (2002). Systems biology: a brief overview. *Science* 295, 1662–1664.
39. Lartigue, C., Glass, J. I., Alperovich, N., Pieper, R., Parmar, P. P., Hutchison, C. A., 3rd, et al. (2007). Genome transplantation in bacteria: changing one species to another. *Science* **317**, 632–638.
40. Gibson, D. G., Benders, G. A., Andrews-Pfannkoch, C., Denisova, E. A., Baden-Tillson, H., Zaveri, J., et al. (2008). Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science* **319**, 1215–1220.
41. Grabowski, M., Joachimiak, A., Otwinowski, Z., and Minor, W. (2007). Structural genomics: keeping up with expanding knowledge of the protein universe. *Curr. Opin. Struct. Biol.* **17**, 347–353.
42. Rappuoli, R., and Covacci, A. (2003). Reverse vaccinology and genomics. *Science* 302, 602.
43. Tettelin, H., Saunders, N. J., Heidelberg, J., Jeffries, A. C., Nelson, K. E., Eisen, J. A., et al. (2000). Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* **287**, 1809–1815.
44. Pizza, M., Scarlato, V., Masignani, V., Giuliani, M. M., Aricò, B., Comanducci, M., et al. (2000). Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* **287**, 1816–1820.
45. Telford, J. L., Margarit, I., Maione, D., Masignani, V., Tettelin, H., Bensi, G. et al., (2004). Vaccines against pathogenic streptococci, in *Genomics, Proteomics and Vaccines* (Grandi, G. , ed.)., Wiley, London, pp. 205–222.
46. Goldschneider, I., Gotschlich, E. C., and Artenstein, M. S. (1969). Human immunity to the meningococcus. I. The role of humoral antibodies. J. Exp. Med. **129**, 1307–1326.
47. Giuliani, M. M., Adu-Bobie, J., Comanducci, M., Aricò, B., Savino, S., Santini, L., et al. (2006). A universal vaccine for serogroup B meningococcus. *Proc. Natl. Acad. Sci. U S A* **103**, 10834–10839.
48. Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome." Proc. *Natl. Acad. Sci. U S A* **102**, 13950–13955.
49. Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. (2005). The microbial pan-genome. *Curr. Opin. Genet. Dev.* 15, 589–594.
50. Maione, D., Margarit, I., Rinaudo, C. D., Masignani, V., Mora, M., Scarselli, M., et al. (2005). Identification of a universal Group B streptococcus vaccine by multiple genome screen. *Science* **309**, 148–150.

Chapter 19

DNA Microarray for Molecular Epidemiology of *Salmonella*

Stephan Huehn and Burkhard Malorny

Abstract

Salmonellosis is a common infection estimated to affect 3 billion people and to cause 200,000 deaths every year. Infections can appear as enteric fever, gastroenteritis, bacteremia, or extraintestinal focal infection. The course of the disease depends on a variety of factors, including infective dose, immune status of the host, and the genetic background of both the host and the pathogen. It has been recognized that certain *Salmonella* types play a major role in the epidemiology of *Salmonella*. Here we describe a DNA microarray comprised of 282 sixty-mer oligonucleotide probes to study the epidemiology of *Salmonella enterica* subsp. *enterica* isolates at the genotypic level. The probes detect targets encoding genes associated with pathogenicity, antibiotic resistance, fimbriae, prophages, flagella (H antigens), lipopolysaccharides (O antigens), plasmids, insertion sequence elements, and metabolism. The probes are printed on glass slides, and whole-genomic fluorescence-labeled *Salmonella* DNA is hybridized to the substrate. For quality assurance, a number of controls are included on the microarray.

Key words: Characterization, epidemiology, microarray, *Salmonella* spp, typing.

1. Introduction

Salmonella is a major zoonotic food-borne pathogen that causes outbreaks and sporadic cases of gastroenteritis worldwide in humans *(1)*. Currently, two species are recognized in the genus *Salmonella*: *Salmonella enterica* and *Salmonella bongori*. *Salmonella enterica* has been subdivided into six subspecies: *S. enterica* subsp. *enterica* (designated subspecies I), *S. enterica* subsp. *salamae* (subspecies II), *S. enterica* subsp. *arizonae* (subspecies IIIa), *S. enterica* subsp. *diarizonae* (subspecies IIIb), *S. enterica*

D.A. Caugant (ed.), *Molecular Epidemiology of Microorganisms*, Methods in Molecular Biology, Vol. 551
DOI: 10.1007/978-1-60327-999-4_19,

subsp. *houtenae* (subspecies IV), and *S. enterica* subsp. *indica* (subspecies VI). Subspecies I strains are usually isolated from humans and warm-blooded animals, while the other subspecies are usually recovered from cold-blooded animals and the environment. According to the Kauffmann-White scheme, subspecies are further divided into serotypes, which are widely used as an epidemiological standardized typing method. Serotyping is based on the antigenic variability of the lipopolysaccharide moieties (O antigens), flagellar proteins (H1 and H2 antigens), and capsular polysaccharides (Vi antigens).

In general, *Salmonella enterica* from human infections can be subdivided into two groups: the enteric fever (typhoidal) group and the nontyphoidal salmonellae, which typically cause gastroenteritis but occasionally, under certain conditions, can cause invasive disease. Mainly, five serotypes are involved in enteric fever: Typhi, Paratyphi A, Paratyphi B, Paratyphi C and related serotypes (Choleraesusis), and Sendai *(2)*. The other approx 2,500 known serotypes belong to the nontyphoidal salmonellae. Although rare, nontyphoidal salmonellae can cause systemic disease, typically when the host's defense is compromised. Specific nontyphoidal serotypes appear to be associated with rather high ratios of invasiveness compared to other serotypes, for example, Dublin, Heidelberg, Brandenburg, and Virchow *(3)*. *Salmonella* Enteritidis and *S.* Typhimurium are the most epidemiological important serotypes because they are responsible for more than 80% of all human infections worldwide *(4)*.

A number of virulence factors and virulence mechanisms have been identified in *Salmonella*. Among those are the type III secretion system, the lipopolysaccharide, and intracellular survival and pathogenicity islands (SPIs) that play a major role in the pathogenicity and epidemiology of *Salmonella*. The various *Salmonella* genomes contain horizontally acquired genetic elements that might play a role in infection, host adaptation, disease development, and spread of antibiotic resistance determinants. Lateral gene transfer is a major contributor to *Salmonella* evolution *(5)*.

A recently identified region that is associated with enhanced virulence is the *Salmonella* Genomic Island 1 (SGI 1), encoding multidrug resistance. SGI 1 is a chromosomally encoded gene cluster, with a size of 43 kb, originally found in a Canadian *S.* Typhimurium phage type DT104 isolate. Recently, SGI 1 was also detected in other, epidemic *Salmonella* serotypes (e.g., Agona, Albany, Newport, Paratyphi dT+) *(6)*. DT104 is associated with enhanced virulence and multidrug resistance. SGI 1 apparently spreads horizontally and represents a public health concern in regard to the future treatment of *Salmonella* infections.

We describe here a protocol to produce a DNA microarray comprised of 282 oligonucleotide probes to study the epidemiology of *Salmonella*. Furthermore, a protocol for the hybridization

and an example to analyze the results are provided. With this DNA microarray it is possible to analyze the presence or absence of a defined gene set of a few hundred specific *Salmonella* target sequences within one experiment (*see* **Appendix**).

2. Materials

All buffers and double-distilled water must be sterilized by either autoclaving or filtration.

2.1. Salmonella *DNA Purification*

1. Microcentrifuge (Eppendorf, Hamburg, Germany).
2. Vortex mixer.
3. Thermal block (thermomixer 5436, Eppendorf) or water bath, capable of being heated to 95°C.
4. DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany).
5. Proteinase K (>600 mAU/mL) solution (Qiagen).
6. Ribonuclease (RNase) A (100 mg/mL) (Qiagen).

2.2. Microarray Production

1. Microarrayer (QArray Mini, Genetix, UK).
2. Thermal block.
3. 384-well polypropylene microarray plate with cover (e.g., Genetix, UK).
4. Pretreated glass slides (CodeLink Activated Slides, GE Healthcare).
5. Desiccator.
6. Slide holder tube (AdvaTube, Advalytix, Munich, Germany).
7. C6 amino-linked oligonucleotides, synthesis scale 0.04 μ*M*, 100 pmol/μL concentration (Metabion, Munich, Germany).
8. 2X print buffer: 300 m*M* sodium phosphate. Dissolve 0.41 g NaH_2PO_4, 3.79 g Na_2HPO_4 in 100 mL distilled water, adjust to pH 8.5.
9. Blocking solution: 100 m*M* Tris-HCl, 50 m*M* ethanolamine, pH 9.0 (*see* **Note 1**).
10. Glass chamber with slide rack (e.g., staining dish with tray, Schiefferdecker type, Duran, Mainz, Germany).

2.3. DNA Labeling and Purification

1. Thermal block or water bath capable of being heated to 95°C.
2. Water bath.
3. Microcentrifuge.
4. Concentrator.

5. Exo-Klenow fragment (5 U/μL, Klenow fragment of DNA polymerase I, GE Healthcare).
6. 10X nucleotide mix with 5-aminohexylacrylamido-dCTP in 10 m*M* Tris-HCl and 1 m*M* ethylenediaminetetraacetic acid (EDTA), pH 8.0 (Bio Prime Plus Array CGH Genomic Labelling System, Invitrogen).
7. Panomer 9 solution covalently labeled at 5′-end with fluorophore AlexaFluor 555 or AlexaFluor 647 (Bio Prime Plus Array CGH Genomic Labelling System, Invitrogen).
8. Reaction buffer 2.5X: 125 m*M* Tris-HCl, 12.5 m*M* $MgCl_2$, pH 7.5.
9. Stop buffer: 0.5*M* EDTA, adjust to pH 8.0.
10. Binding buffer B2: 40% (v/v) 2-propanol, supplied with Labelling System (Bio Prime Plus Array CGH Genomic Labelling System, Invitrogen).
11. Washing buffer W1: 80% (v/v) ethanol, supplied with Labelling System (Bio Prime Plus Array CGH Genomic Labelling System, Invitrogen).
12. Elution buffer E1: 10 m*M* Tris-HCl, pH 8.5.
13. DNA purification spin columns with collection tubes (Bio Prime Plus Array CGH Genomic Labelling System, Invitrogen).
14. 1.5-mL amber reaction tubes (Eppendorf).

2.4. Hybridization

1. Hybridization chamber (sciHYBCHAMBER, Scienion, Berlin, Germany).
2. BfR hybridization buffer: 400 μL formamide, 100 μL Denhardt's solution (Fluka), 100 μL 10% (w/v) sodium dodecyl sulfate (SDS), 150 μL 20X SSC (3*M* NaCl, 0.3*M* sodium citrate, pH 7.0), 250 μL 20% (w/v) dextrane sulfate.
3. Water bath heated to 42°C.
4. Lifter slips (MSeries 22 × 26.5 mm Erie Scientific Company, Portsmouth, NH).
5. Wash solution I (1X SSC + 0.3% w/v SDS): 150 m*M* NaCl, 15 m*M* sodium citrate, and 1 m*M* SDS in double-distilled water.
6. Wash solution II (0.2X SSC): 30 m*M* NaCl, 3.0 m*M* sodium citrate in double-distilled water.
7. Wash solution III (0.05X SSC): 7.5 m*M* NaCl, 0.75 m*M* sodium citrate in double-distilled water.

2.5. Scanning and Data Analysis

1. Two-color laser scanner (excitation by 532 and 635 nm), including analysis software (e.g., GenePix 4000B scanner, Gene Pix Pro 6.0 software, Axon Instruments, CA).

2. Table calculation software (e.g., Microsoft Excel, Redmond, WA).
3. Visualization and analysis software (e.g., BioNumerics 5.0, Applied Maths, Ghent, Belgium).

3. Methods

The microarray protocol described consists of the following steps: (1) *Salmonella* DNA purification from pure cultures; (2) production of the microarray; (3) genomic DNA labeling; (4) microarray hybridization of labeled genomic DNA; and (5) scanning and data analysis.

3.1. Salmonella *DNA Purification*

For the purification of *Salmonella* DNA, the DNeasy Blood and Tissue Kit from Qiagen is used.

1. Transfer 1.6 mL *Salmonella* overnight culture (16–18 h incubated at 37°C) into a clean 2.0-mL reaction tube. Centrifuge at 10,000*g* for 4 min (*see* **Note 2**).
2. Discard the supernatant carefully.
3. Resuspend the pellet completely in 180 µL ATL buffer by vortexing.
4. Add 25 µL proteinase K; mix by vortexing briefly.
5. Incubate the suspension at 56°C for 3 h using a thermomixer at 750 rpm.
6. Centrifuge the tube for 10 s.
7. Let the tube cool at room temperature to approx 40°C.
8. Add 5 µL RNase A (100 mg/mL), mix by vortexing, and incubate for 5 min at room temperature.
9. Mix by vortexing for 15 s. Add 210 µL AL buffer and mix thoroughly by vortexing.
10. Add 210 µL 96–100% ethanol and mix by vortexing immediately to yield a homogeneous solution.
11. Pipet the mixture into the DNeasy Mini spin column placed in a 2-mL collection tube.
12. Centrifuge at 10,000*g* for 1 min. Discard the collection tube and place the column in a new collection tube.
13. Add 500 µL AW1 buffer and centrifuge at 10,000*g* for 1 min. Discard the flowthrough and collection tube and place the column in a new collection tube.
14. Add 500 µL AW2 buffer and centrifuge at 10,000*g* for 1 min.

15. To dry the DNeasy membrane, centrifuge for 3 min at 17,500*g*. Discard the flowthrough and the collection tube.
16. Place the DNeasy Mini spin column in a 1.5-mL microcentrifuge tube and pipet 50 µL AE directly onto the DNeasy membrane.
17. Incubate at room temperature for 5 min, then centrifuge at 10,000*g* for 1 min to elute DNA.
18. Repeat elution step and centrifuge at 17,500*g* for 2 min.
19. Store the DNA at 4°C until fluorescence labeling.

3.2. DNA Microarray Production

The source plate should be prepared in a room free of *Salmonella* DNA (*see* **Note 2**).

3.2.1. Preparation of the Source Plate

1. Dilute 6 mL of print buffer with 2.4 mL double-distilled water. Fill the 384-well microarray plate with 21 µL per well using a multichannel pipet. Add 9 µL of each 100 m*M* oligonucleotide probe using a multichannel pipet (*see* **Note 3** and **Appendix**).
2. Add 9 µL of double-distilled water in the wells that do not contain oligonucleotide probes.

3.2.2. Print Process

Here we describe the printing conditions and application of a QArray Mini microarrayer (Genetix) for the array production. For other microarrayers, the printing conditions should be adapted.

1. Place the slides onto the slide holder. The activated surface must be placed on top (*see* **Note 4**).
2. Place eight pins in the print head and adjust the head for approx 0.5-mm inking depth. The source inking order is set by rows.
3. For slide design, select 8-pins/7-fields order and arraying by fields. For field layout, use fields 1, 2, 5, and 6 (*see* **Note 5** and **Fig. 1a**).
4. Set the pattern dimension to 160 µm estimated spot size, row count 6, column count 8, row pitch 750 µm, column pitch 500 µm.
5. Set the number of blots required before printing on the sample slides to 5 and the blot pitch to 650 µm.
6. Set the washing program between oligonucleotide inking to 3,000 ms washing using distilled water, 500 ms waiting. Repeat washing step six times. The final step is 3,000 ms washing, 35,000 ms drying with compressed air, and 5,000 ms waiting (*see* **Note 6**).

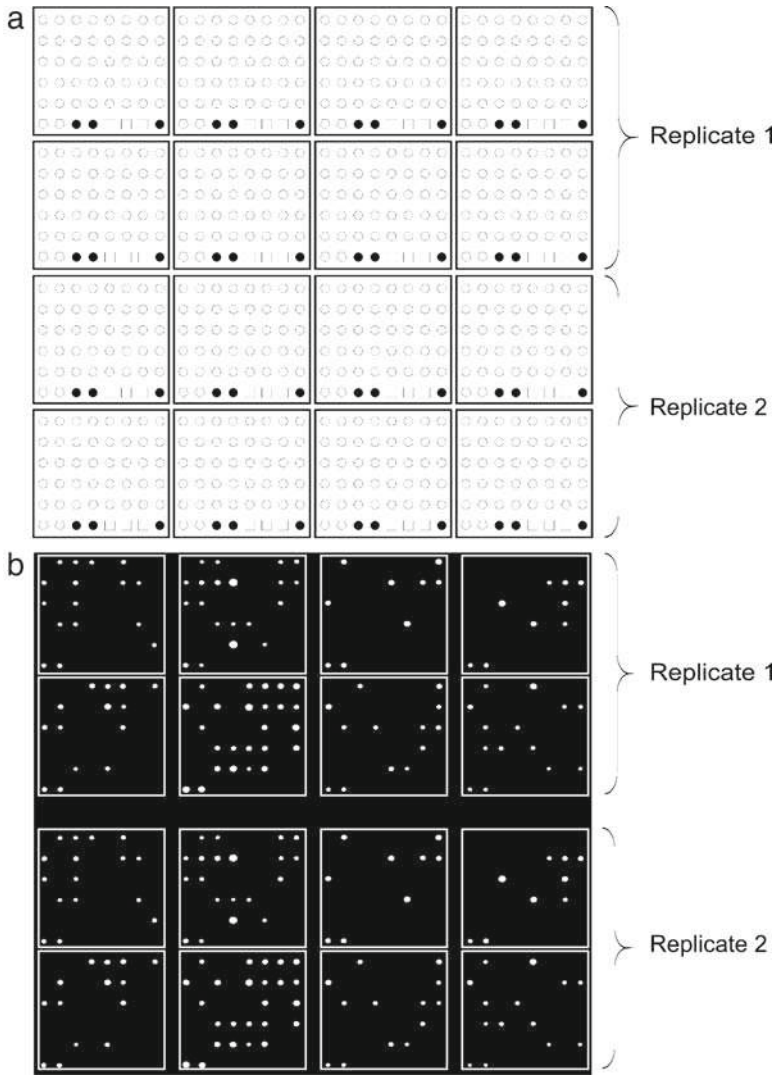


Fig. 1. *Salmonella* DNA microarray. (**a**) Scheme of the array showing the segmentation into 16 blocks. All blocks contain a control row including positive control spots (diamond), print buffer (contamination) control spots (black circle), and negative control spots (square). The oligonucleotide probes are positioned within the upper 5 rows of each block (white circle). Each probe is printed as replicate. (**b**) Hybridized *Salmonella* DNA labeled with AlexaFluor 555. A gray spot indicates the presence of a target sequence within the genome.

7. Print the slides using 24 stamps per inking. Set the stamp time to 10 ms and inking time to 2 s.
8. After printing, place the slides immediately in a sealed chamber (e.g., glass chamber with slide rack) containing saturated sodium chloride solution at the bottom. Close the

chamber with a lid and seal the chamber with parafilm. Keep the chamber overnight at room temperature.

3.2.3. Postcoupling Processing

1. Block residual reactive groups using preheated blocking solution at 50°C for 30 min with gentle shaking in a sealed chamber.
2. Rinse the slides in distilled water and afterward in postprint washing solution.
3. Wash the slides using preheated postprint washing solution at 50°C for 30 min with gentle shaking in a sealed chamber.
4. Rinse the slides in distilled water.
5. Wash the slides using preheated distilled water at 50°C for 30 min with gentle shaking in a sealed chamber.
6. Place four slides each in a slide holder tube and spin the tube in a centrifuge for 3 min at 5,000*g*.
7. Store the slides until use at room temperature, protected from light and humidity.

3.3. Labeling of Genomic DNA

For the labeling of the genomic *Salmonella* DNA, the BioPrime Plus Array CGH Genomic Labeling System is used. General requirements for the labeling according to standard laboratory praxis have to be considered (*see* **Notes 2** and **7**).

3.3.1. DNA Labeling

1. Pipet approx 4 μg *Salmonella* genomic DNA in a maximum volume of 24 μL in an amber 1.5-mL microcentrifuge tube and adjust the volume to 24 μL with sterile water (*see* **Note 8**).
2. Add 20 μL fluorophore-random oligonucleotide mix (Panomer 9 resuspended in reaction buffer), mix by vortexing gently, and briefly centrifuge to the collect the contents.
3. Incubate the tube at 95°C for 10 min, immediately cool on ice, protected from light, for 5 min.
4. On ice, add 5 μL 10X fluorophore nucleotide mix with AlexaFluor 555-aha-dCTP or AlexaFluor 647-aha-dCTP and 1.5 μL exo-Klenow fragment (*see* **Note 9**). Mix gently briefly by vortexing and centrifuge to collect the contents.
5. Incubate the tube at 37°C for 3.5 h in a water bath protected from light.
6. Add 5 μL stop buffer to the tube and place on ice.

3.3.2. Purification of Labeled DNA

1. Add 200 μL binding buffer B2 to the labeled DNA and briefly mix by vortexing.

2. Load the sample onto the PureLink Spin Column placed in a 2-mL collection tube.
3. Centrifuge at 10,000*g* for 1 min. Discard the flowthrough and place the column in a new collection tube.
4. Add 650 µL of wash buffer W1 onto the column.
5. Centrifuge at 10,000*g* for 1 min. Discard the flowthrough and place the column in a new collection tube.
6. Centrifuge the column at 17,500*g* for 3 min to remove any residual wash buffer.
7. Place the column in a new sterile amber 1.5-mL collection tube.
8. Add 55 µL elution buffer E1 to the center of the column and incubate at room temperature for 5 min.
9. Centrifuge the column at 17,500*g* for 2 min. The flow-through contains the purified labeled DNA.
10. Dry the eluate in a vacuum concentrator at 60°C for 25 min.

3.4. Microarray Hybridization of the Labeled DNA

Protect the labeled DNA from light as much as possible at all steps (*see* **Note 2**).

3.4.1. Hybridization

1. Place the slide containing the printed probes into the hybridization chamber. Place one lifter slip per array field onto the slide. Fill 30 µL of sterile water into humidity wells and prewarm the closed chamber at 42°C for 10 min.
2. Add 30 µL of hybridization buffer to the labeled and dried DNA and resuspend the DNA by careful pipeting. Avoid air bubbles.
3. Incubate the sample at 95°C for 2 min.
4. Centrifuge briefly to collect the content and to avoid liquid at the tube walls.
5. Open the prepared hybridization chamber.
6. Load the sample carefully under the lifter slip. Avoid air bubbles.
7. Close the hybridization chamber and incubate at 42°C for approx 18 h in a water bath.

3.4.2. Washing

1. After incubation, open the hybridization chamber and remove the slide from the chamber with tweezers. Immediately rinse

the slides in 300 mL wash solution I, preheated at 34°C, to remove the lifter slips.

2. Place the slide in a slide holder and wash it at 34°C for 3 min in wash solution I with gentle shaking.
3. Rinse the slide with wash solution II, preheated at 34°C.
4. Wash the slide at 34°C for 3 min in wash solution II with gentle shaking.
5. Rinse the slide with wash solution III, preheated at 34°C.
6. Wash the slide at 34°C for 3 min in wash solution III with gentle shaking.
7. Rinse the slide in sterile water at room temperature.
8. Place the slide in a holder tube and centrifuge the tube at 5,000*g* for 3 min.
9. Store the slides until scanning, dry and protected from light, at room temperature.

3.5. Scanning and Data Analysis

1. Prescan the slide using a microarray scanner according to the manufacturer's instructions. Select the green laser light channel for the DNA, if labeled with AlexaFluor 555 and the red laser light for the DNA, if labeled with AlexaFluor 647. Adjust the PMT gain (*see* **Note 10**).
2. Define the array field and perform a full scan with high resolution (10 μm pixel resolution).
3. Save the scanned array in TIFF format (*see* **Fig. 1b**).
4. Automatically align signals for identifying and analyzing individual features using a GAL (GenePix Array List) file (*see* **Note 11**) and quantify feature intensities using the GenePix Pro software. Subtract the local background intensity from each feature intensity. Save the raw feature intensities as a text file and import the file in a table calculation program (e.g., Excel) for normalization.

3.5.1. Normalization

1. Calculate the signal intensity average of the two positive control spots (ttrC probe) of each block. Calculate a ratio between the ttrC intensity and each feature intensity of the corresponding block. Based on the ratio the presence, absence, or uncertainty of the target will be defined. Using CodeLink Activated slides the cutoff ratio is set to 0.25. A normalized ratio over 0.25 is considered as target sequence presence. A normalized ratio between 0.25 and 0.15 is considered an uncertain result. In this case, an individual decision has to be made (*see* **Note 12**). A normalized ratio below 0.15 is considered as target sequence absent.

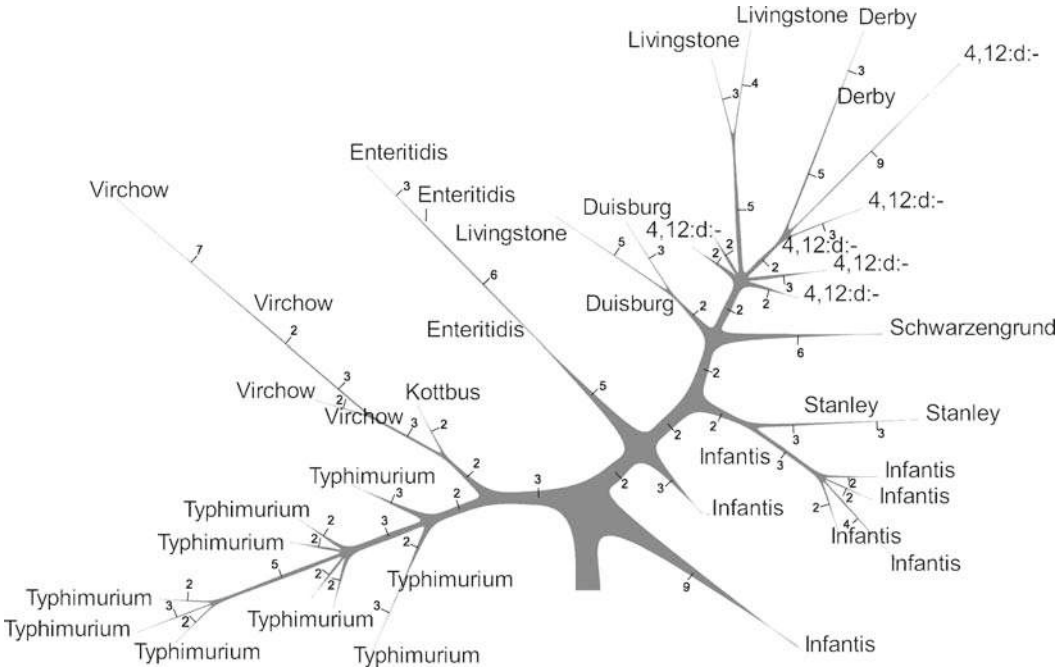


Fig. 2. Maximum parsimony tree calculated by BioNumerics 5.1. The tree exemplarily shows the differences of various *Salmonella* isolates tested based on 75 pathogenicity genes. Strains belonging to a certain serovar cluster together. The small numbers indicate the number of genes different between the branches.

2. Import a table indicating the presence or absence of each individual target from one strain to BioNumerics 5.1. Select certain target marker groups (e.g., pathogenicity) and perform a maximum parsimony tree to visualize the differences in the gene set between the strains tested (*see* **Fig. 2**).

4. Notes

1. Prepare the solution without adding ethanolamine; adjust to pH 9. Add ethanolamine directly before use.
2. Use pipet tips with filters only. Carefully avoid cross-contamination of the samples. The reagents, especially the enzymes, should be kept on ice during pipeting. To avoid fluids on the reaction tube wall, spin the tubes shortly in a microcentrifuge before use.

3. To be able to use multichannel pipets, order oligonucleotide probes in a 96-deep well microtiter plate.
4. The print chamber and slide surfaces should be free of dust. The humidity should be between 25 and 50%.
5. One field is divided into eight blocks (eight columns, six rows) representing a full set of probes (see Fig. 1a). All fields have the same probe assignment. Consequently, one slide contains two arrays; fields 1 and 2 form the first array, fields 5 and 6 the second array.
6. The compressed air should be totally clean. Avoid propellants and oil aerosols. Especially, propellants influence the surface tension, which influences the spot size and may generate extremely large spots.
7. After adding the fluorophores carefully protect the solutions from light for the whole process. Especially, UV light will bleach and lower the signal intensity. Use amber reaction tubes.
8. At least 4 μg of genomic DNA should be used, but not more than 10 μg.
9. Generally, AlexaFluor 555-aha-dCTP can be used for labeling. We have observed that this fluorophore labels DNA more efficiently, resulting in stronger signal intensities, than using AlexaFluor 647-aha-dCTP. There is no difference in the specificity of the fluorophores.
10. Positive control spots (ttrC probe) should be used for defining optimal signal intensity. Pixels with intensities out of the range (e.g., over 65,000 raw intensity) shall be strongly avoided since it is not an accurate measurement of the pixel intensity.
11. A GAL (GenePix Array List) file defines an array of blocks to match the size and positioning of printed features and to apply substance names to the features. The file can be usually generated by the microarrayer software.
12. The signal intensity depends on the DNA quality, the hybridization reaction, as well as the labeling reaction. Moreover, the cutoff value may differ between different chips with different surfaces and fluorophores.

Acknowledgments

We thank Cornelia Bunge-Croissant and Ernst Junker for technical assistance.

Table 1
Appendix. DNA Microarray Probe Sequences and Their Specific Characteristics

Accession No.	Probe name	Marker group	Gene function	Probe sequence	Length
X01385	aac(3)-IV	Resistance	Aminoglycoside-3″-acetyltransferase, encoding GEN resistance	GACACGATGCCAACACGACGCTGCATCTT-GCCGAGTTGATGGCAAAGGTTCCCTATG	57
AJ310480	aacC1	Resistance	Aminoglycoside-3″-acetyltransferase, encoding GEN resistance	CTTATGTGATCTACGTGCAAGCAGATTACGGT-GACGATCCCGCAGTGGCTCTCTATA	57
S68058	aacC2,3	Resistance	Aminoglycoside-3″-acetyltransferase, encoding GEN resistance	GAAGAAACGGTGAAGTCGCCTGGAAAACGGCAT-CAGAATACGATTCAAACGGCATTC	57
AJ009820	aadA1a	Resistance	Aminoglycoside-3-adenyltransferase, encoding STR/SPE resistance	GAAGTGGTGATCGCCGAAGTATCGACTCAAC-TATCAGAGGTAGTTGGCGTCATCGAG	57
AF261825	aadA2,3,8	Resistance	Aminoglycoside-3-adenyltransferase, encoding STR/SPE resistance	AAATTTCGAACCAACTATCAGAGGT-GCTAAGCGTCATTGAGCGCCATCTGGAATCAA	57
AF169041	aadA5,4	Resistance	Aminoglycoside-3-adenyltransferase, encoding STR/SPE resistance	GTTCTTGCTCTTGCTCGCATTTGGTACAGCGCT-TCAACTGGTCTCATTGCTCCTAAG	57
AF078527	aadB	Resistance	Aminoglycoside-2″-adenyltransferase, encoding GEN resistance	CATGGAGGAGTTGGACTATGGATTCTTAGCGGA-GATCGGGGATGAGTTACTTGACTG	57
AE008792	abe_B	Serotyping	CDP-abequose synthase: Serogroup B	ACCTTCATATACTGAGTATCAAGTTGGAACTGGT-GCTGGGGTAAGTTTGAAAGATTTTCTGGT	63
X61917	abe_C2-C3	Serotyping	CDP-abequose synthase: Serogroup C_2-C_3	TGCATTAAAGCGTCCTATAACCGAGCCAACGAT-TATCAATACCTTGATTGAATGGTTGA	59
NC_003197	acrF	Resistance	Acridine resistance protein F	GACATCTCTGACTATGTCGCCTCTAACATTAAG-GATTCTATCAGCCGTCTGAATGGT	57
U43280	agfA	Fimbrial	Thin aggregative fimbriae precursor	GAACTGACTCAGAATGGTTTCAGAAATAAT-GCCACCATCGACCAGTGGAACGCTAAA	57

(continued)

Table 1 (continued)

Accession No.	Probe name	Marker group	Gene function	Probe sequence	Length
AF024666	aphA1-Iab	Resistance	Aminoglycoside-3″-phosphotransferase, encoding KAN resistance	TTTGACGAGGGGAAATTAATAGGTTGTATTGATGTTGGACGAGTCGGAATCGCAGAC	57
DQ177329	armA	Resistance	16S rRNA methylase, extended-spectrum β-lactamase- resistant	AAAGTCTTTATCTGGAAAGGAGAAGGGAATGGAAGAGAATTACCAGCTATGGTTTGAATCTTT	63
AF013573	avrA	Pathogenicity	SP1- encoded protein; inhibits the key pro-inflammatory	ACCGAAGCATTGACCTGTATTGTTGAGCGTCTGGAAAGTGAAATTATAGATGGCAGC	57
AE008835	barA	Pathogenicity	Sensory histidine kinase	GCGGCTCGACCTTCTGGTTTCATATTAATCTTGATCTTAACCCAAATGTCATTATTGACGGGC	63
AE008694	bcfC	Fimbrial	Fimbrial usher, bovine colonization factor	GCACAGTCAGGAACCAATTTACAGCTTATGGGCTATCGCTATTCAACCTCGGGCTTT	57
AJ238349	$bla_{oxa1\ like}$	Resistance	Extended- spectrum β-lactamase, encoding AMP resistance	TTCTCTGGAGATAAAGAAAGAAACAACGGATTAACAGAAGCATGGCTCGAAAGTAGCT	58
AJ238349	$bla_{oxa1\ like}$	Resistance	Extended- spectrum β-lactamase, encoding AMP resistance	CCCAAAGGAATGGAGATCTGGAACAGCAATCATACACCAAAGACGTGGATGCAATTT	57
AF153200	bla_{pse1}	Resistance	Extended- spectrum β-lactamase, encoding AMP resistance	AGTATTACAGCAGTTGTGTGGAGTGAGCATCAAGCCCCAATTATTGTGAGCATCTATCT	59
AF153200	bla_{pse1}	Resistance	Extended- spectrum β-lactamase, encoding AMP resistance	GCAAGTTGAACAAGACGTTAAGGCAATTGAAGTTTCTCTTTCTGCTCGTATAGGTGTTTCCG	62
AF309824	$bla_{tem\text{-}1\ like}$	Resistance	Extended- spectrum β-lactamase, encoding AMP resistance	TAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGG	57
AF309824	$bla_{tem\text{-}1\ like}$	Resistance	Extended- spectrum β-lactamase, encoding AMP resistance	AGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGATC	57

AY123253	catA1	Resistance	Chloramphenicol acetyl transferase, encoding CHL resistance	ACATATATTCGCAAGATGTGGCGTGTTACGGT-GAAAACCTGGCCTATTTCCCTAAAG	57
AL627271	cdtB	Pathogenicity	Cytolethal distending toxin, secreted protein	GGAATCTTCAGGGCTCTTCAGCATCTACA-GAAAGTAAATGGAATGTCAATGTCAGAC	57
M64556	cmlA1like	Resistance	Chloramphenicol exporter, encoding CHL resistance	AAATATGGGCTTTGCAGTCCGTGTTAGGCTT-TATTGCTCCAATGTGGCTAGTGGGTATT	59
X77455	cmy-1 (bla fox-1)	Resistance	CMY-type extended- spectrum β-lactamase, encoding AMP resistance	GAGTTCAGAAGAGCTAAGAAGTTGCTTGAGG-TACTGGGTTGCATTGATAATAGTCATG	58
U77414	cmy-2 like	Resistance	CMY-type extended- spectrum β-lactamase: AMP resistance	CTATTCCGGGTATGGCCGTTGCCGTTATCTAC-CAGGGAAAACCCTATTATTTCACCT	57
NC_003197	copR	Resistance	SPI 5, Ccopper resistance protein	CAGGAACATTATTCATTGATTATTCTTGATATTAT-GCTGCCGGGGCTTGATGGATGG	57
AE016840	csgA	Fimbrial	Major curlin subunit precursor	TAGGCCAGGGTGCGGATAACAGTACTATT-GAACTGACTCAGAATGGTTTCAGAAACAATG	60
AE008724	cstA	Metabolism	Carbon starvation protein	ATGAATAAATCAGGGAAATACCTCGTCT-GGACAGCGCTCTCAGTATTGGGTGCGTTT	57
X92507	ctx-M2	Resistance	CTX-M2 extended- spectrum β-lactamase: AMP resistance	TATAGCGACAATACTGCCATGAATAAGCTGATT-GCCCATCTGGGTGGTCCCGATAAA	57
AY341107	cutF	Metabolism	Putative copper homeostasis protein	CAGCCTGTACGCTATTTGCATTGATAGGGTG-TAATAACCGTGCGGAAGTTGACGCCC	57
AY103456	dfrA1 like	Resistance	Dihydrofolate reductase, encoding TMP resistance	ACCCAACCGAAAGTATGCGGTCGTAACACGT-TCAAGTTTTACATCTGACAATGAGAAC	58
AF175203	dfrA12	Resistance	Dihydrofolate reductase, encoding TMP resistance	CGGCAAGCCTCTACCGAACCGTCACACATTGG-TAATCTCACGCCAAGCTAACTACCG	57
AJ313522	dfrA14	Resistance	Dihydrofolate reductase, encoding TMP resistance	GGTCGTTACCCGCTCAGGTTGGACATCAAAT-GATGACAATGTAGTTGTATTTCAGTC	57

(continued)

Table 1 (continued)

Accession No.	Probe name	Marker group	Gene function	Probe sequence	Length
AF220757	dfrA17,7	Resistance	Dihydrofolate reductase, encoding TMP resistance	TCTTCCAAATCGCAAATATGCAGTAGTGTCAAA-GAACGGAATTTCAAGCTCAAATGAA	58
AE008878	dgoA	Metabolism	Galactonate dehydratase	GAAAATAACTCACATCACCACGTACCGTTTAC-CTCCACGTTGGATGTTCCTGAAAATCG	59
AE008723	entF	Pathogenicity	Enterobactin synthetase, component F (non-ribosomal peptide synthetase)	CGCTATTTGGCCCGGTGCTCAACATAAAAGT-GTTTGATTATCATCTGGATCTTCCTGG	58
NC_003197	envR	Pathogenicity	Transcriptional repressor, envelope Rregulator for envCD, acrEF	ACTTATTCAGGATAGGCTTACGGGATGCT-GGAATGATAATCCTTTACAGGATCTACG	57
Y14067	fhuA_Spa	Pathogenicity	Outer membrane protein receptor	GTTACGACTGGGCCGATCAAGAGTCTCTTAAC-CGCACTACTGGCATCACATCTAAAC	57
AE008703	fhuA_STM	Pathogenicity	Outer membrane protein receptor	GTTGACAACGAGCGTTTACAGAATTTCAGCGTA-GATACACAACTGGAAAGTAAATTCGC	59
AE008721	fimA	Fimbrial	Major type 1 subunit fimbrin (pilin)	GACAATAGCACTACCGCAACCGGCGTCGGGATT-GAGATTCTTGATAATACCTCTTCA	57
AE008787	fliC/fljB	Serotyping	Filament structural protein, detects all H antigens	CACAAGTCATTAATACAAACAGCCTGTCGCTGTT-GACCCAGAATAACCTGAACAAATCCC	60
AY649698	fliC_b	Serotyping	Filament structural protein, detects b antigen	ACAGTTACTGAAAACCAAATTGTAGACGCTGTTA-CACCGACGCCAGTTGATACAGTC	57
AL627272	fliC_d	Serotyping	Filament structural protein, detects d antigen	AACCAAATTGCTGAAGTAACAAAAGAGGGTGTT-GATACGACCACAGTTGCGGCTCAA	57
AJ292284	fliC_e,h	Serotyping	Filament structural protein, detects e,h antigen	CTTGAAGCCGGTGGCAAGTACTATGCTGCAAC-CTATGACGAAGGTACAGGTAAAATC	57
AJ292278	fliC_e,n,x-e,n,z15	Serotyping	Filament structural protein, detects e,n,x and e,n,z_{15} antigens	CCCAACTAAATCTACTGTTACAGGTGATACCGCT-GTTACTAAGGTACAGGTTAATGCTCCT	61

M84980	fliC_g,x	Serotyping	Filament structural protein, detects g complex- associated antigens	GTACCGCTGAAGCCAAAGCGATAGCTGGT-GCCATTAAAGGTGGTAAGGAAGGAGATA	57
AE008787	fliC_i	Serotyping	Filament structural protein, detects i antigen	GGTCTTGGTGGTACTGACCAGAAAATTGAT-GGCGATTTAAAATTTGATGATACGACTGGA	60
U06201	flic_m,t	Serotyping	Filament structural protein, detects m,t antigen	CAACTCAGGGGCGGTAGTAACTGACACCACT-GCTCCAACTGTTCCTGATAAAGTATA	57
X04505	fliC_r	Serotyping	Filament structural protein, detects r antigen	AAGTCACTTTAACTGGCACACCAACAGGAC-CAATTACTGCTGGCTTCCCTTCAACTG	57
AY434692	fliC_z10	Serotyping	Filament structural protein, detects z10 antigen	AAACTGCTGGAATTACTGGTGCTACAT-TAAAAGCTGGTATTACTGGTACAACGACA-GAAACCG	63
AY649736	fliC_z4,z23	Serotyping	Filament structural protein, detects z4,z23 antigen	AAATTAGATGTGACTAAAGGAATCGCAACCACTG-TAAGCTCTGGAGCCTCGGTAGTT	57
AE008826	fljA	Serotyping	Repressor of phase-1 flagellin	GTGTGAGGACATCCAATGGCAATCATAT-GCAAGTTATGTTTGACTGGGTGAGCAGGA	57
AE008826	fljB_1,x	Serotyping	Filament structural protein, detects 1,x antigens	CCAATAATGGTACTACACTGGATGTATCG-GGTCTTGATGATGCAGCTATTAAAGCGGCT	59
AJ292277	fljB_l,w-l,v	Serotyping	Filament structural protein, detects l,w and l,v antigens	ACAGTGGTATTAGTGCTGCTGATGCTGCAAAAG-GTCAATTAGTTACGATGTCTTATACGGA	61
AF118107	floR	Resistance	Putative efflux protein: FLO/CHL resistance	CTGATGGCTCCTTTCGACATCCTCGCTTCACT-GGCGATGGATATTTATCTCCCTGTC	57
AF246666	gipA	Pathogenicity	Gifsy-1 Peyer's patch-specific virulence factor	CTGATTAACGATAACCAAGTTGTATGCGTC-GAATCCCTGAAAGTGAGGAACATGATC	57
AE008720	glxK	Metabolism	Glycerate kinase II	GAAACGGGTTTCTGTGGACGTTAGCGGGCCGAT-GGGGGAAAAAGTAAACGGATTTTAT	58
AE008818	gogB	Pathogenicity	Gifsy-1, leucine-rich repeat protein	TGGGACAGGAAGAATAGAGCCGTGTTTAATAAA-GATGAGAAGATAGCAGAAAGATTGAATGA	62

(continued)

Table 1
(continued)

Accession No.	Probe name	Marker group	Gene function	Probe sequence	Length
AF254762	gtgA	Pathogenicity	Gifsy-2 prophage protein	TTCCAGACCTTCCAGAACACCAAGATAATCCT-TCGCAATTACGCCTCCAACATGATG	57
AL513383	HCM1.71	Mobility	Putative periplasmic protein, on pHCM1 plasmid (R27)	CCATGTAATTTCAATATGTACGCCAAGTTAGGT-CAAAGACTGGTGGGATCTCGGGAA	57
AE008831	hilA	Pathogenicity	SPI 1- encoded transcription activator	TGATGATTTCATACTCAACATGGACGGCTCCCT-GCTACGCTCAGAAAAGAAAGTCAATA	59
AE008902	hilD	Pathogenicity	Putative AraC-type DNA binding domain containing protein	ACTTTTCGGCCCCATTAACAAAACCGAC-GACAAAACATCTGTTAGCGCCAATAGAAA	57
AE008826	hin	Serotyping	Regulator for fljA: DNA invertase hin	TACATGAACGTGGAGCTCACTTCCATTCTTTAAC-CGATAGTATTGATACCAGTAGCGCG	59
AY462995	hldD_ DT104	Pathogenicity	DT104- specific phage- encoded protein	AAAGGTCAATGACCATTGTTCTGTTCATCGCAT-AGGTTTCAGCAGACTCTATAAGCG	57
AE008913	hsdM	Metabolism	DNA methylase M, host modification	AAAACTACGTCAATGAACTCGCCTCGCTGCT-GTTTTTGAAAATGTGCAAAGAGACCG	57
AY144490	htrE	Pathogenicity	Probable porin/fimbrial assembly protein	TTGTGCGTGGTATTAACAATGCTGGTGAACT-CATCGTTCGTTGGTATGAAGAAGGTC	57
L16014	hydH(stn)	Pathogenicity	Enterotoxin sensory kinase	ATTCAGGGAGTGAGTAATAATATCATTGAGGT-TAACCGTCTGGAGCGTCAGATGCGC	57
AE008831	iagB	Pathogenicity	SPI 1- encoded invasion associated protein IAGB precursor	CATAACCGAGATGGTTCAACCGATCTTGGCCT-GATGCAAATTAACAGCTTCCATATG	57
AF261825	int_SG1	Resistance	Integrase from Tn*4555*, present in *S.* Typhimurium DT104 SGI 1	CTATCTCTACGAGAACCCAAAGACACAAGCA-GAGCGTCAGCACAATAAAGAAATGTTGC	59
X12870	int1	Resistance	DNA Integrase 1: Iintegron associated	GATCTGCTCGGCCATTCCGACGTCTCTACGAC-GATGATTTACACGCATGTGCTGAAAG	58

L10818	int2	Resistance	DNA Integrase 2: Integron associated	GCAAGAACTCTTAGGGCATAACGATGTTAAGAC-CACGCAAATCTATACGCATGTGTTGG	59
M90846	invA	Pathogenicity	SPI 1- encoded invasion protein	TGTTCGTCATTCCATTACCTACCTATCTGGTT-GATTTCCTGATCGCACTGAATATCGTACTGG	63
AE008832	invH	Pathogenicity	SPI 1- encoded invasion protein	AACCCGGAAGTAAGAATTTAAGCATATATCA-GACGTTACTTGCTGCCCATGAAAGACTGCAA	62
AE008832	invI	Pathogenicity	SPI 1- encoded secreted protein	CAGACAGCTCAGTCGTGAGGAAATTTATACGT-TATTACGTAAGCAGTCTATTGTTCGCCGG	61
AE008826	iroB	Pathogenicity	Putative glycosyl transferase	GTCGTTGGACCACTGATTGCCGCTAAGTATGA-CATTCCGGTAGTGATGCAAACCGTC	57
AY328029	irsA	Pathogenicity	Putative transcriptional regulator, internal stress response element, prophage encoded	GGTCTATAAGGCTGTGCTGAACGTCCTGT-GGAACTGGATTCTGTTTCGTAAATTTTC	57
X07037	IS150	Mobility	Insertion element, unnamed protein ORF B	TTGTCTGGATAATGCTGTGGTGGAGTGTT-TCTTTGGAACCTTAAAGTCGGAGTGTTT	57
AJ746361	ISCR1	Mobility	Insertion sequence common region ISCR1	GAAGATCCGAAGGTCATTGAGCAGATTCTCAAG-CATCTGAAACAGAAAACAGCCAAG	58
AF231986	ISCR2	Mobility	Insertion sequence common region ISCR2	GATCGGCGCTCAATCTGAATGTTCACTTCCA-CATGCTGTTTCTCGACGGTGTGTATG	57
AF261825	ISCR3	Mobility	Insertion sequence common region ISCR3	CGACGACAGCATGGATGGGCTGCGGATGAGT-TCGATCACCTACC	57
AY341249	ISCR4	Mobility	Insertion sequence common region ISCR4	CTGCCATGCGGAACGATTGGTGGCGTTCTCGT-GCAAGAAG	44
AY114142	ISCR6	Mobility	Insertion sequence common region ISCR6	GAACTTCCTATACCCTTCTCCTGTACCT-GCGAATTGTGCCATGCTGATACGTCGAAA	40
AJ250371	ISCR7	Mobility	Insertion sequence common region ISCR7	TGAACGAGCATGTGCATTTCCATTGCTGTGT-CATCGACGG	47

(continued)

Table 1
(continued)

Accession No.	Probe name	Marker group	Gene function	Probe sequence	Length
AF028594	ISCR8	Mobility	Insertion sequence common region ISCR8	GCAAAGAGGGAAGCAAACTGGTGTACCGCTGT-GCCAAACA	40
AF106956	rep_iterons_ FIA	Mobility	F Plasmid pmk115, mini F plasmid replication origin, IncFIA replicon	CATGACGGTATCTGCGAGATCCATGTT-GCTAAATATGCTGAAATATTCGGATTGACC	40
AE008868	rep_iterons_ HI2	Mobility	Plasmid R478, incompatibility group HI2 (IncHI2 subgroup)	GACTTAATAGGCTCACTACCGTTGTCATCCTG-TAAGTTAAGAGGTTGATCTGCTCAA	57
AF106566	rep_ iterons_P IncP	Mobility	Plasmid RK2, transposon insertion site (Tn*1723*) plasmid incompatibility group	CACAGATGATGTGGACAAGCCTGGGGATAAGT-GCCCTGCGGTATTGACACTTGAGGG	57
K03089	leuO	Pathogenicity	Transcription regulator, component of ilvIH-leuO-leuABCP promoter relay region	GCTGGGCGTGATAAAGGGCATCAATGGATGGAA-GATTTATTAGTCTCTGTTTGTAAGCGATA	57
AF106566	lpfD	Fimbrial	Long polar fimbrial operon protein	GGTGGAACCTATGCGATGTCCTGTGAATGCCCT-GATGATACCTCTCTTATAAATGAC	62
AE008875	marT	Pathogenicity	SPI 3- encoded putative transcriptional regulator	AAAGTGTGGGAAGAAGACGGCATGGTGGT-GTCGGCAAATACGCTTTATCAGAATATC	57
AF161317	merA	Resistance	Hg(II) reductase: Mercury resistance	ATGAGCACTCTCAAAATCACCGGCATGACTT-GCGACTCGTGCGCAGTGCATGTCAAG	57
AE008725	mgtC	Pathogenicity	SPI 3- encoded putative transcriptional regulator	CAGTGGTTACTGAATATCGTAAAAGAG-GCCGCGATCTGTTTACAAGGGTTAGGTTCG	57
AE008799	misL	Pathogenicity	SPI 3- encoded protein	AACCGATTATTCCTGATCCAGTAGACCCTGT-TATCCCTGACCCTGTCGTTCCCGATC	57

NC_003197	mntH	Metabolism	NRAMP manganese transport protein	GATAATTGTCGTGCTGGTCGTCGCCCT-GAATATCTGGTTGCTGGTTGGAACGGTGAT	57
AE008800	mrdA	Metabolism	Cell elongation- specific transpeptidase of PBP 2	GATCGCCTGTCAGAATGGATGGGTAAGTTCG-GCTATGGTCACTACACCGGAATTGAT	57
X60666	msgA	Pathogenicity	SsrB-regulated factor	GATATCATTCTGGGAGAATTGACTAAGAGGGT-TCACCGGATCTTCCCCGATGCTGAT	57
X60665	nanH	Resistance	Fels-1- encoded neuraminidases	TGGTCATTGCCATCAGGATACTGTGAGGGTTTT-GGTTCAGAAAATAACATCATAGAA	57
AF378725	oafA	Metabolism	Acetylation of the O:5- antigen	GCACATCTTATGCCAGGGCTGAAATCGGTATTT-GGAAATTCACTTAACATTACGCAGAGAAC	57
AE008831	orf 17.4	Serotyping	β- O-Polysaccharide polymerase: Sserogroups E_1 and E_4	TCACTTTATCACTAAATGGTCCGTTCCTGTCTA-CATTGCATTAGGTTTGCTACTGCCTCT	62
AE016845	orf 9.6	Serotyping	α- O-Polysaccharide polymerase: Sserogroup E_1	AGGTGGGCAGTTTATATTCACAGAGGCTTT-TCATTCACTTGGTTATGTCGGAGTATTCCTG	60
AE008792	orf84	Phages	Putative prophage Cro protein	CCGCCTCATGTGAACGATAGCAGAACCATATT-AGCGAAGGTGAACAACCATGATCAC	61
AF013776	orgA	Pathogenicity	SPI 1- encoded oxygen-regulated invasion protein	TTCACCTGAACAATTTGTTGAAACGGCAGTAG-GCGTCATTAAGCATCATCTTGATGAACTTC	57
AE008783	oxyR	Pathogenicity	Hydrogen peroxide-inducible regulon activator	GCTGCATATTGGTTTAATCCCAACAATTGGTC-CCTATTTGCTGCCGCTTATCATTCC	62
AF250878	pagA(udg)	Metabolism	UDG-glucose/GDP-mannose dehydrogenase	GGTGGCGTTAGACATTGTTCCTTCCCGTGTT-GAACTGTTAAATGATCGGATATCTCC	57
AB041905	pagJ	Pathogenicity	Gifsy-3 encoded, PhoP-activated gene	GGCTTTAATATTACCATCCTCACTATATCTT-GCTCTTACAATAGCGGCAGACTCCCAA	57
AE008892	pagK	Pathogenicity	PhoPQ-activated protein	TTTGGCCTGCTGGCATACCCTTCCCTGAAGATT-GGTTTAAAATGTGTAGAGGTTATTG	58

(continued)

Table 1
(continued)

Accession No.	Probe name	Marker group	Gene function	Probe sequence	Length
AE008753	parA-parB IncHI1	Mobility	Putative partition protein on antibiotic resistance plasmid	GTTCCCGATATGCTGCTCATCACTCATACATTAC-CCCCTTGCTTCACTGAAAAACTC	59
AE008753	pefA	Fimbrial	Virulence plasmid- encoded fimbrial protein	AACCAGGTTGTTCAGTTAGGTACTGTTCAG-GCAGGTCAGGAAGGTACGGCTGTTGATT	57
NC_006431	pflD	Metabolism	Putative pyruvate formate lyase II	ATGACCCATCGTATTCAACGCCT-CAAAGCCGCTCTGTTTCAGAATCACCGT-GAAATT	58
NC_004631	phoP	Pathogenicity	Response regulator in two-component regulator system with phoQ	GAATACACCATTATGGAAACGCTTATCCG-TAACAACGGTAAAGTGGTCAGCAAAGATTCGC	57
AE008747	phoQ	Pathogenicity	Sensory kinase protein in two-component regulatory system with PhoP	TGATGGGCAACGTACTGGACAACGCTTG-TAAATATTGTCTGGAGTTTGTCGAGATTTC	61
AY532917	pilR	Pathogenicity	SPI 7- encoded nucleotide-binding protein, putative sigma 54 interaction protein	AATGTGGTTATGCCTTCCCCTCAAAA-GAAGCCGTGAACTATCTCTCTATGGTCAGTG	58
NC_003197	pilV	Pathogenicity	SPI 7- encoded prepilin peptidase	CAAACCAATCGTTTCACTCAGGCAGTCTCATC-CTATGTTGGAAAGTTTTATCCGACG	57
NC_003197	pipA	Pathogenicity	SPI 5- encoded protein	AGAAGGCAGGAAAGTTATTTGTCTCAATCT-GGACGATTCTGATGATTCATATACCGAACA	57
U66901	pipB2	Pathogenicity	T3SS translocated protein	GCTGGACAAGTTATTGTACGAGTCAGTAAAG-GCGACCATTCTGAGACAAGAGAAATTCCG	60
AF261825	pipD	Pathogenicity	SPI 5- encoded protein	CAAACAGCTAAGCAGCAGTATAAGATGGAGCA-GAGCTATCTGAGATTATATGCGTCG	57
AY906856	prgH	Pathogenicity	SPI 1, needle complex inner membrane protein	AAAGAGAAGACGATAACAAGCCCAGGGCCATA-CATAGTTCGATTACTTAACAGCTCA	57

AM234698	Prot6E	Fimbrial	*Salmonella* Enteritidis fimbrial biosynthesis protein	TGTGGGTCGTAACGCACAAGTGAAATTTACCCTGAGGGAGGCTTATGGTAATAATAATTTGG	57
AM234722	qacEΔ1	Resistance	Qac multidrug exporter, encoding Et-Br and quaternary ammonium resistance	GCAATAGTTGGCGAAGTAATCGCAACATCCGCATTAAAATCTAGCGAGGGCTTTACTAA	62
NC_003197	qnrA	Resistance	Confers quinolone resistance, plasmid located	ATGGATATTATTGATAAAGTTTTTCAGCAAGAGGATTTCTCACGCCAGGATTTGAGC	59
AE006471	qnrB2_B1	Resistance	Confers quinolone resistance, recogniszes both *qnrB1* and *qnrB2*, plasmid located	CAGCAAACTTCACACATTGCGATCTGACCAATTCGGAGTTGGGTGACTTAGATATTC	57
AE008837	qnrS	Resistance	Confers quinolone resistance, plasmid located	ATGGAGAGGGTTTGTTTAGAAAAATGTGAGTTGTTTGAAAATCGCTGGATAGGAACG	57
J01724	ratB	Pathogenicity	Putative outer membrane protein	CTATTTATCTAAATCAACCTCAAGCGGCGGCATGTATTGCGGCGTAGATGAGAATAC	57
BX664015	rck	Pathogenicity	Resistance to complement killing protein, encoded on pSLT plasmid	GTACAGTTTAATCCGGTGGAAAATGTGGTCATCGATCTGGGCTATGAGGGAAGTAAAGT	59
M20134	recC	Mobility	DNA replication protein	CTCATGGGGAAAACTGGGGCGGGATTATATCTACCTGCTATCTGAACTGGAAAACAG	61
Y00768	rep_ori_γ IncX	Mobility	IncX replicon - plasmid incompatibility group	GTTAGCCATGAGGGTTTAGTTCGTTAAACATGAGAGCTTAGTACGTTAAACATGAGAGC	57
M20413	rep_SG1	Pathogenicity	Replication protein encoded by SGI 1	TTCAAGAGTAAATGGATCTGAGTTTGAGAGGCGTTGTGGTGCTGCTGGTCTATACAT	59
M28718	rep_W	Mobility	Plasmid R388, class 1 integron In3, plasmid incompatibility group IncW	GTCTTGATGATCTCGTTGATACGATGGCCGGGGGCTTTGTTGTTCTTAGGCATGTTG	57
M93063	repA_A/C	Mobility	Replicon A, plasmid incompatibility group	GGACCACCAGCTAGAAAGTATTAACGGAACAATCATGAGTAAGAGAACCAAAGACAAA	57

(continued)

Table 1 (continued)

Accession No.	Probe name	Marker group	Gene function	Probe sequence	Length
AF261825	repA_B_C IncL/M	Mobility	Plasmid pMU407.1, plasmid incompatibility group, Inc L/M	TATACTCAGGTGGTTATAGTCGTTTGTAGTCGGAGGGCTTGTGAGCAGAGTAGTTGAG	58
U12441	repA_FIB	Mobility	IncF plasmid RepFIB replicon, plasmid incompatibility group	GAAGTAAGTTAATGACATAAACTATGGTCAGTATGCCAGACTCAGTTGTTAAATACAGGCTGC	58
X73674	repA_FIIs	Mobility	pSLT, plasmid incompatibility group	CATTATGATCCACTGGCCAACCGTGTACAGAGATCCATCACCAATCTGGCTATAGAG	63
U27345	repA_N	Mobility	IncN plasmid R64, plasmid incompatibility group	TATCTGGGAAATCGAATTTAACCATAAACTCCTGCGGTACATTTACGGCCTGACGAA	57
M26308	repA_repB IncF	Mobility	Plasmid pRK100, plasmid incompatibility group	GTACTGCGAGAGAGAGGGGATAACACAGGCTCAGTTCGTTGAGAAAATCATCAAAGAT	57
AE006471	repA_T	Mobility	Plasmid Rts1, plasmid incompatibility group IncT	TAATCAAAGTAGTATAACTCCCATAATCGCTCGTCTGCTTCCAGTTCCACAACGTCTGTATCG	58
NC_003292	repA_Y	Mobility	IncY plasmid P1, plasmid incompatibility group	TCTTTACGCAGACATTGAAAGTAAGGCAAAAGAACTAACAGTTAATTCAAACAACACT	63
AY234375	repA2_FIC	Mobility	Plasmid F, putative replication protein genes	ATCTTCACATTGATTCCAGCAAGTATCCTCACCCGTTTTGCAGCCTTCTCCAGAAAA	58
K00053	repC_DT193	Mobility	DNA replication protein	AAGCACATACAGGACTGCATCGAGCGCCTTTGGAAGGTATCCATCATCGCCCAGAAT	57
K02380	repC_R64	Mobility	DNA replication protein of plasmid R64	GGAAAATACTTCGGTCAACGCCCTAGCCGAACGCTTCCTCGATGACGGCCTGAAAAC	57
M16168	rfbD	Serotyping	TDP-dehydrorhamnose synthetase: Sserogroups A, B, C_2-C_3, D_1, D_2	TGACCTTATTCTGCCTCAATGGGAATTAGGAGTTAAGCGTATGCTGACTGAAATGTT	57

AY524415	rfbE	Serotyping	CDP-tyvelose-2-epimerase: Serogroups A and D	ACAGTGGTGTTCAGGCATTCATCAATGTATGGT-GGGAGACAGTTTGCTACTTATGATC	60
AP005147	rhaA	Metabolism	L-rhamnose isomerase	GTCTGATCCCAGGCCCAAAACGGCTGAACCTA-CACGCCATTTACCTTGAGTCGGATA	57
AE008792	rhuM	Pathogenicity	SPI 3- encoded putative cytoplasmic protein	GAAGCCGAAGGTGAGAAGGATATCGCCGGTTT-GCTACAATGGGAAACAGAACCTAAA	58
AL627273	rep_RNAI IncI1	Mobility	Plasmid incl-1 mini-replicon, plasmid incompatibility group	CATAAGCGACAGCTTGTGGCAGGTCTGAA-GAATACTCCATATAACGCAGTACACTGG	57
AE008889	rep_RNAI_ BO	Mobility	Plasmid pMU720, IncB mini-replicon encoding RNA I, RNA II, and pro-motor regions	TCACATAAGGATGTATCTGTGGCAAGAGCGAA-GATAAGCAGTTGAATAGATCGTTATATT	57
AE008874	rep_RNAI_ K/B	Mobility	Plasmid R387, replication-associated protein, plasmid incompatibility group	CAGCTTGTGGCAGGTCTGAAGAATACTTCAT-ATAACGCAGTACACTGGAGTCAGTTAGC	57
AE008833	rpoS	Metabolism	Major sigma factor during stationary phase	GTCTGCTGGGATATGAAGCTGCGACACTGGAA-GATGTAGGCCGTGAAATCGGTCTTA	60
AE008708	safC	Fimbrial	Putative fimbriae usher protein	GTAAGTGCTAGTTGGCAGATGACTTCACCAT-CACACGGTGGTCAGACGCAACAAGTG	59
M63169	sat (Tn*7*)	Resistance	Streptothricin acetyltransferase of Tn*7*	AAGAGCTTGTCGGGAAGATTGAACTCAACTCAA-CATGGAACGATCTAGCCTCTATCG	57
AB161461	sat1 (int2)	Resistance	Streptothricin acetyltransferase linked to integrase 2	CAAGCTATGAGCCAGGTCAAGCTCCATATTC-CGTTGAAGAATTAGCAGATGATGTGG	57
NC_004313	SB10	Phages	Encoded by ST64B phage	GTTATGTGCTCTCCCGCGAATTGCGAAAAT-CACTCCGTCCTCAGCGAAGGAAGAAAA	57
NC_004313	SB54	Phages	Encoded by ST64B phage	TTGAACTACTGGCAATCAGATATGCAGCATGGAT-TAAGCCGGAATTTGAAATCGAAG	57
L11008	sefA	Fimbrial	Fimbrial protein encoded by *S.* Enteritidis, *S.* Dublin and *S.* Gallinarum	TTTCCGTGGGCGTATTCAGGGAGCCAATAT-TAATGACCAAGCAAATACTGGAATTGA	57

(continued)

Table 1 (continued)

Accession No.	Probe name	Marker group	Gene function	Probe sequence	Length
AF239978	sefR	Fimbrial	Sef14 fimbrial regulator	AGCCGTTGTGGAAATTGAATATATGGATGACATT-GAATCATTTGACATTATTACTTTGCCAGA	57
Xxxxxx[1]	SEN4287	Metabolism	Putative restriction endonuclease gene *Salmonella* Enteritidis unique PT4	GTCGGTAGTGATTAGCGTACCGCGAATGGCTTT-GTGGTAGGGTAGTGCACCGCGTAG	57
AE008719	sfbA	Pathogenicity	Putative periplasmic iron-binding lipoprotein	CGACGCCAACCTCTTTCAACATACCCTCTATTTT-GACAAATTCACCGCTGACAAAGG	63
NC_003197	sgbE	Metabolism	Ribulose-5-phosphate 4-epimerase	GAAATTAACGGCGAGTACGAGTATCAGACCG-GCGAGGTGATTATTAAAACCTTTGAA	57
NC_003197	shdA	Pathogenicity	Fecal shedding factor	CAATATGGTTTCTACAACAACAGCGTAGA-GAGCGGTGATGCGGGATCTGAATCTTAT	57
U51867	sifA	Pathogenicity	Lysosomal glycoprotein (lgp)-containing structures	CAAGAAAAGGCAACCTACCTGGCAGCGAAAAT-TCAGTCTGGGATTGAAAAGACAACG	57
AE008831	sipA	Pathogenicity	SPI 1- encoded *Salmonella* (cell) invasion protein	GCAGTAACCATAGCGTGGATAACAGTAAGCATAT-TAACAATAGCCGAAGCCATGTCG	57
AF026035	sirA	Pathogenicity	Invasol SirA: Regulator of invasion proteins	GTAAAGGTGGGGTACGTAATGATGACACATATC-CGAATGCCAGTAACAATGCCGAAG	57
AF128999	sitA	Pathogenicity	Invasion SirA: Regulator of invasion proteins	GATATTAAACGAGCGCAGGGGGCACAGCTTATC-CTCGCGAATGGTCTGAACCTGGAG	57
AF127079	slrP	Pathogenicity	Leucine-rich repeat protein	TATGATAACAGCATAAGGACACTGCCAGCA-CATCTTCCGTCAGAGATTACCCATTTGAATG	57
AE008762	slyA	Pathogenicity	MarR family transcriptional regulator for hemolysin	GAGCTTCTGATTAAACTTATCGCCAAACTT-GAACACAATATTATGGAATTGCACTCTCACGA	57
AF007380	sodC1	Pathogenicity	Gifsy-2- encoded, copper/zinc superoxide dismutase	TATCCGTTACTGGCACCACGCCTTAAATCACT-GTCAGAACTGAAAGGTCACTCATTG	61

AF254764	sodCIII	Pathogenicity	Fels-1- encoded putative Cu/Zn superoxide dismutase precursor	TATTACAATTACCGAAACAGAATATGGCTTGT-TATTCACCCCACATTTGTCCTCACTTCC	62
AF121227	sopA	Pathogenicity	Secreted outer protein	CAGATAATTTCCTGCTGCTTTCCTCCCAAGAT-TCAGACACGGCGATGATGCTCTCCA	57
AE008747	sopB	Pathogenicity	SPI 5 invasion gene D protein	CGGCAAAGATCGTACAGGGATGATGGATTCA-GAAATCAAGCGAGAGATCATTTCCTTACATC	60
AE008834	sopD	Pathogenicity	Secreted outer protein	TGCCCGGCTCATCAAGATCTGTTTACTATCAA-GATGGACGCTTCTCAGACACAATTT	57
AL627268	sopD2	Pathogenicity	Secreted effector protein, sopD homologue (pseudogene)	TGGAATGTGTTGAATGGAATGGTACGCTTACT-GAAGAGGAGATGAATAAACTACGCTGTCT	63
L78932	sopE1	Pathogenicity	Translocated effector protein, encoded by P2-like cryptic bacteriophage	ACGTTTTATTTCGCATAAGAACACTGAATCTTCT-GCAACACACTTTCACCGAGGAAGC	57
AF200952	sopE2	Pathogenicity	Secreted outer protein	GTGACTAACATAACACTATCCACCCAGCACTACA-GAATCCATAGAAGTGACGTTGAACCAGT	61
AE008832	spaS	Pathogenicity	SPI 1- encoded surface presentation antigen	ATATTGTAGGTATTGCCGTCATTTGGCGT-GAACTTCTCCTCGCATTGGTATTAACTTGC	58
AF060869	siiD (spi4_D)	Pathogenicity	SPI 4 HlyD family secretion protein, predicted cation efflux pump	ATCTCTTTCTAAAGGAGGGACGATACAAGATATT-TATGTAGCCGAGGGTGATACTGT	62
AF060869	siiE (spi4_F)	Pathogenicity	SPI 4- encoded protein	TAATGGTATTGCTGTCGGTCAGGCTGTAACGGA-TAGTTTGGGTAACTTCACCTTTAC	59
AF060869	siiF (spi4_R)	Pathogenicity	SPI 4- encoded putative ABC-type bacteriocin/lantibiotic exporter	GGGCGTGAGTTATCAGTATGATGCTCAATCTC-CGATGATTATTAACCGACTGTCTAT	57
NC_003197	sprB	Pathogenicity	SPI 1- encoded transcriptional regulator for Type III Secretion System (TTSS)	TGTGTGCTGCAATATTTTGGCGTTATGGATTAT-GTTTTAAAAGACGAACTATCCTGCG	57
AE008831	sptP	Pathogenicity	SPI 1- encoded tyrosine phosphatase	ACCTTATTAAGAGCAAGGATAATGTTGGTGT-CAGGAATGCCGCTTTAGTCATAAAAGGC	57

(continued)

Table 1
(continued)

Accession No.	Probe name	Marker group	Gene function	Probe sequence	Length
AE006471	spvC_a	Pathogenicity	*Salmonella* plasmid virulence: hydrophilic protein	GGTTACGATGTTTTCATCCATGCTCGTCGA-GAATCACCTCAGTCTCAGGGCAAATTT	58
AE006471	spvC_b	Pathogenicity	*Salmonella* plasmid virulence: hydrophilic protein	CGGTGACAAGTTCCACATCAGTGTGCTCAGGGA-TATGGTGCCACAAGCATTTCAAGC	59
AE006471	spvR	Pathogenicity	*Salmonella* plasmid virulence: regulation of spv operon, lysR family	ACAAAGACTCTTTATACGGAAGAATGGCACTCT-TATCCCAACCGAATTTGCACAAACT	57
AE008907	srfJ	Pathogenicity	Putative virulence factor, activated by transcription factor SsrB	TGACATGATTGGTAATTTCAAATCGGGTTG-TAGCGGGTTTATCGACTGGAATCTGCTG	57
AE008761	ssaQ	Pathogenicity	SPI 2- encoded secretion system apparatus protein	ATACCACAACAGGTGCTCTTTGAGGTCGGACGT-GCGAGTCTGGAAATTGGACAATTA	58
NC_003197	sseC	Pathogenicity	SPI 2- encoded translocation machinery component, required for systemic infection	AAATAGAGCAATTAATAACTCAGCAACGGTTTCT-GGATTTCATAATGCAACAAACAGA	58
AE008761	sseF	Pathogenicity	SPI 2- encoded secretion system effector	ATTCATATTCCGTCAGCGGCAAGTAATATAGTC-GATGGTAATAGTCCTCCTTCCGATATA	57
AE008743	sseI	Pathogenicity	Gifsy-2 prophage putative type III secreted protein	GCAACAGAACCGGGAGTGGAACGCACAGA-TATAACTTACAACCTAACCAGTGATATTGAT	58
AF294582	sseJ	Pathogenicity	Translocated effector protein regulated by SPI 2	CACATCATATCTTACCCTCCTATGGT-CAATACTTTGGCGGAAGGTTTACTAATGGATT-TACC	60
AE008894	sseK1	Pathogenicity	T3SS *Salmonella* secreted effector K1 secreted by SPI 2	GGGATAATAGCTGTTGATCGCAATAACCACCCG-GCTTTACTTGCTGGATTAGAAATAATGC	60

AE008795	sseK2	Pathogenicity	*Salmonella* secreted effector K1	TTTCATGTCAAAGTAATACTCAAACCATCGCAC-CTACGCTCAGTCCACCTTCATCAG	62
AF013776	sspH1	Pathogenicity	Gifsy-3- encoded leucine-rich repeat protein, *Salmonella*-secreted protein H1	CTTACCTTCCCCGCGTGGGAGGAGAATAT-TCAGTGTAACAGGGATGGTATAAATCAG	61
AE008800	sspH2	Pathogenicity	Leucine reach repeat protein, *Salmonella*-secreted protein H2	GATGTCTTCCCGCCACCATCAGTAATCGCCG-CATTTATCGTATTGCCTGGTCTGATA	57
AE008761	ssrB	Pathogenicity	SPI 2- encoded protein: secretion system regulator	CCTGTTGTGCATACGAGCCTGACATACTTATC-CTTGATCTTAGTCTACCTGGCATCA	57
AL627265	staA	Fimbrial	Fimbrial protein encoded by *S.* Typhi CT18	CGGCTGATGTAACTGATGCCACTAAGGCT-TCTCTGGTAATGGATTCGTCATTCTCTA	57
AE008710	stbD	Fimbrial	Fimbrial usher protein	AATATCGGTTTGCCAACGGTGATTAGCGTCAG-TAATAGTGAAACATTTCAACCCTCCG	57
AE008795	stcC	Fimbrial	Paral putative outer membrane protein	TGTAGTAGATCATCATGGTCATAATGTGGGCATT-GTTGGACAAGGTAGTCAGCTATTTATTCG	57
AE008839	stdB	Fimbrial	Putative outer membrane usher protein	AATTACTGGAACGCACAGTCCAACAATAACTA-CATGCTCAGCCTCAACAAGGTGTTC	58
AL627276	steB	Fimbrial	Outer membrane fimbrial usher protein encoded by *S.* Typhi CT18	CAGCCCGGATCAGAGTAACTATAACCT-GTCTCTTTCCTGGTACTTCGACTTAGGGTC	63
AE008703	stfE	Fimbrial	Putative minor fimbrial subunit	GGCGGTGAGGTGGAATTTGGCAATGTGTTGAC-GACGAAAGTGGATGGGGTGAATTAC	57
AL627280	stgA	Fimbrial	Fimbrial protein encoded by *S.* Typhi CT18	GATTCTGCGTATAGCACGATTGATACCACAGCG-GGTACGGCTTCTATGGAGTTTATC	57
AE008702	stiC	Fimbrial	Putative fimbrial usher protein	TGAACTACAGCTTCAGCGGCTATAAAGAGTACG-GTTCCAGTGAGGATTCCGACGATG	57
AE008915	stjB	Fimbrial	Putative fimbrial usher protein	GGTTATTACACTTATCAGGCTACGGATAAT-GACAACGACTCTCGCAGTATAATGGCTTCCT	57

(continued)

Table 1
(continued)

Accession No.	Probe name	Marker group	Gene function	Probe sequence	Length
NC_006511	stkC	Pathogenicity	Outer membrane usher protein	CGGAACCTATCGTGTGAAAGTGAATCTTAACAAT-GCGTTGAAATCTACCTCTGAAATTACC	57
AE0098709	STM0305	Serotyping	Discrimination between *Salmonella* subspecies I and non- subspecies I	CACCTGTTCAAAGATTTTCTGAAGGCGCAGGAG-TATTCATTACTGATATCCTCCATT	61
AE008710	STM0330	Metabolism	Putative 3-isopropylmalate isomerase, (dehydratase), subunit with LeuC	TGCGAAATTACGACGGCAGAGGAAACGATCTC-CTTTGTGATCAGTGAACTCAAACGG	61
AE008737	STM0900	Phages	Fels-1 prophage- encoded protein	GTAAGCAACTACATCAGACACAAACTTATCAGA-TATGCAAAAGGAAAGGGGAAGGCGGC	57
AE008784	STM1896	Metabolism	Putative cytoplasmic protein	TTTCAGTAGATGTTTCCGACAATGGTATTTCT-GGCGTGGCAAAAGAGTGGCTTAAAG	57
AE008819	STM2616	Pathogenicity	Gifsy-1 prophage- encoded protein	ACCACTCAAATCTCTGTCGAAACTCTTTC-CCCGATTACCCATAACCAAATTCCCGTTA	59
AE008823	STM2701	Pathogenicity	Fels-2 prophage- encoded protein	GAATGCTGATCTGGCCTGACTTCATCAACTTT-GACACCGTGCTGAAAGCAGACGCGA	57
AE008824	STM2740	Pathogenicity	Fels-2 prophage- encoded protein	CTACAGAACGCTTCTATCCGCAAGGAAAC-CCGCTGCCTTATCGAATGGGAGCTACTG	58
AE008842	STM3098	Serotyping	Discrimination between *Salmonella* and non-*Salmonella*	GCTATGGGAAGACCAGATTATCTATATTTAT-CACTCTACGCCGGGTTCGCAGGGAAG	57
AE008876	STM3782	Metabolism	Putative PTS system galactitol-specific enzyme IIC component	AGGTAACCCAGCCATTATATCTACAGCACTGAT-TCTGACACCTATCTCTGTCTTTATTGC	57
AE008889-1	STM4057	Serotyping	Discrimination between *Salmonella* subspecies I and non- subspecies I	TGATCATTACGTTGTGATTTATTCCAACGG-TACGCTGTATGGGGAATGGCCC	57

AE008896	STM4200	Phages	Putative phage tail fiber protein H	TGGAGGTGGAGGGCATACGAGTAATACAGAT-GGTCTTCTCTATTGTTCAGGTGGTAA	60
AE008896	STM4210	Phages	Putative methyl-accepting chemotaxis prophage- encoded protein	GATGAACTGCTGAGCGTCGTTGAAGAGGGTAT-GCGTGAAGCCAAAGAGATGATGGAT	52
AE008911	STM4497	Metabolism	Putative cytoplasmic protein	AAAAACAACGGCTCCGGTAATGAGATTGGGT-TCTGGATTTTTGATTATCCTGCTCAG	57
AE008916	STM4595	Fimbrial	Putative fimbrial chaperone protein	GATTACGTTCAATGGCAAAATTTACGATCAG-GCGTGTACGGTTCAGGTGAATGGCTC	57
M28829	strA	Resistance	Aminoglycoside-6 ″-phosphotransferase, encoding STR resistance	GAACAGCAGATCGCTATGCCGATTTGGCACT-CATGATTGCTAACGCCGAAGAGAACT	57
M28829	strB	Resistance	Aminoglycoside-3 ″-phosphotransferase, encoding STR resistance	GGACTCCTGCAATCGTCAAGGGATTGAAAC-CTATAGAAGACATTGCTGATGAACTGC	57
AL627279	STY3672	Phages	Hypothetical phage protein encoded by phage cs73 of *S.* Typhi CT18	ACAGAAGATTCCATTACAGATGTGTTTAAC-CAAGTGCCTTCATCGTGTATATCCGACAGTGC	57
AL627279	STY3676	Phages	Phage protein, putative capsid scaffold-ing protein	CTGATTAACCTTGAGCACATCAAGTCTTATCT-GCCGGACAGCACCTTTAACCGCTAC	57
AL627281	STY4221_1	Metabolism	Putative amino-transferase	ATTAATGGATTTCCAGTAATACCTATGTCTGTT-GCAGATACCGATTTAGTTTCTCCCGTGG	62
AL627283	STY4625	Phages	Phage protein, major capsid in phage P2, homologue with Fels-2 protein	CGATCCGACGCTGATGGAAGACGTGGAATA-CAAATGCGAGCAGACCAACTTTGATAC	57
AL627283	STY4631	Phages	Hypothetical phage protein encoded by phage cs73 of *S.* Typhi CT18	ACATAAGCCTGAAGGTAAGAAGCAGAAG-CACACGCATCCTGTAGATGTTGTATTTAGT	61
AF106566	sugR	Pathogenicity	SPI 3- encoded putative ATP- binding protein	CGCATTTCCACTAATCCAGTTTATTGTCACTAC-CCATAGCCCGCAGGTTATCAGCAC	57

(continued)

Table 1
(continued)

Accession No.	Probe name	Marker group	Gene function	Probe sequence	Length
X12869	sul1	Resistance	Dihydropteroate synthase encoding SUL resistance	CTCTTAGACGCCCTGTCCGATCAGATGCACCGT-GTTTCAATCGACAGCTTCCAACCG	58
M36657	sul2	Resistance	Dihydropteroate synthase encoding SUL resistance	TTCTATCCGCAATTGGCGAAATCATCT-GCCAAACTCGTCGTTATGCATTCGGTGCAA	57
AJ459418	sul3	Resistance	Dihydropteroate synthase encoding SUL resistance	AAATAACTGGAACCGATGTGAAATCTCGTTTAG-CACCAACTCTTGCAGCAGAAATGTATGC	57
AL627266	tcfA	Fimbrial	Typhi coloniszation factor, putative fimbrial protein	GTATGCCACAGGAGAAGGAGGTACCAGCAG-GGAATGATATAGAGACAGGACTTGTTG	57
X61367	tet(A)	Resistance	Efflux pump, encoding TET resistance	ATCGTCGGACCCCTCCTCTTCACGGCGATCTAT-GCGGCTTCTATAACAACGTGGAAC	61
V00611	tet(B)	Resistance	Efflux pump, encoding TET resistance	TTGGATGGAATAGCATGATGGTTGGCTTTTCATT-AGCGGGTCTTGGTCTTTTACACT	57
J01749	tet(C)	Resistance	Efflux pump, encoding TET resistance	CATGACTATCGTCGCCGCACTTATGACTGTCT-TCTTTATCATGCAACTCGTAGGACA	57
L06798	tet(D)	Resistance	Efflux pump, encoding TET resistance	CGGAGCAGAAACAAGAAAGCGCAGGTAT-CAGCTTTATCACACTGCTTAAACCTCTGG	57
L06940	tet(E)	Resistance	Efflux pump, encoding TET resistance	CGGCGTTATTACGGGAGTTTGTTGGAAAG-GCTAATGTTGCAGAGAACTACGGTGTTT	57
S52437	tet(G)	Resistance	Efflux pump, encoding TET resistance	GCCTCACCAATCTAAGCTCTATCGCAGGAC-CGCTTGGCTTCACAGCACTCTATTCTG	57
NC_006816	tnp_Cf	Mobility	Transposase insertion sequence IS	CATCGACCAGTGCTTGGATCTCAGTGATC-TACGTGCCTACCTGGCAGATTTCTATAG	57
NC_006816	tnp_IS1	Mobility	Transposase insertion sequence IS*1*	GGTGGAGCTGCATGACAAAGTCATCGGGCAT-TATCTGAACATAAAACACTATCAATA	57

NC_002056	tnp_IS102	Mobility	Transposase insertion sequence IS*102*	GTGAATATGCAGACCGTAACCGTGCAGT-GGCTAATCAGCGAATGACCGGGAGTAATG	57
NC_003384	tnp_IS1202	Mobility	Transposase insertion sequence IS*1202*	CTATTTTGAAGCGACCCGTGGCTATATCGA-GAAATACGGTAAGCCCATGATCCTTTA	57
NC_006816	tnp_IS1294	Mobility	Transposase insertion sequence IS*1294*	TTTTAAGATGGTGAGGTACTTCGGGTTCCTT-GCCAACCGTGTGTGTGGAGAGAAGCT	57
NC_003198	tnp_IS1351-like	Mobility	Transposase insertion sequence IS*1351*like	GAATCACAAAAGGCTCCACCGTATTTACTGTCT-GCTCAAGCTGAATTTTCGCCGTAA	57
NC_002305	tnp_IS30	Mobility	Transposase insertion sequence IS*30*	GCTAAACAACAGACCGAGAAAGACACTGAAGT-TCAAAACACCGAAAGAGATAATTGA	57
AJ310778	tnp_Orf341E	Mobility	Transposase Orf341E	CTCAATGTCCACTACCACATGCTGTTTCTCGAT-GGTGTCTATGCCGAAGATGACTAT	57
AJ634602	tnp_pFPTB1	Mobility	Transposase insertion sequence on plasmid pFPTB1	TACTTTGGTAATAACAGAGGGATCACCTGGTA-CAACTTTGTGTCCGATCAGTATTCC	57
NC_006511	tnp_SPA2465	Mobility	Transposase SPA2465	GCTGAATGAGGTGCGGGAAATTACGGATAAAG-GGTTATCAGAATATAACTGCGAACG	57
NC_003198	tnp_STY343	Mobility	Transposase STY343	CAGATCATCGCTGTGATTAGATCAGTTGAATC-CGGACGGACTGTTAAAGATGTCTAC	57
NC_003384	tnp_Tn2680	Mobility	Transposase of Tn*2680*	CTTTGAATGGGTTCATGTGCAGCTCCATCAG-CAAAAGGGGATGATAAGTTTATCACC	57
NC_003384	tnpA_IS1-like	Mobility	Transposase A of insertion sequence IS1 like	TGCAGTTCACTTACACCGCTTCTCAACCCGG-TACGCACCAGAAAATCATTGATATGG	57
AJ746361	tnpA_IS1696	Mobility	Transposase A of insertion sequence IS*1696*	TCATATCAACCCGAAGTATGGAAGCAGTCCAG-GACGGACTTTCTACACCCATATCTC	57
NC_006816	tnpA_IS186	Mobility	Transposase A of insertion sequence IS*186*	CGCTGAATGGCGACTACATATGGGATATGATC-CTCATACCTGTCAGTTCACTGATTT	57

(continued)

Table 1
(continued)

Accession No.	Probe name	Marker group	Gene function	Probe sequence	Length
AY341107	tnpA_IS200	Mobility	Transposase A of insertion sequence IS*200*	AAATACCGAAGACAAGCGTTCTATGGAGA-GAAGCGTAGGGCAGTAGGCAGCATATTAA	57
NC_006816	tnpA_IS26	Mobility	Transposase A of insertion sequence IS*26*	TTGAACACCGACAGATTAAGTACCG-GAACAACGTGATTGAATGCGATCATGGCAAAC	57
AY509004	tnpA_IS3/IS911-like	Mobility	Transposase A of insertion sequence IS*3*/IS*911* like	GAGGAAGAGAACACCAGACTCAAGAAGCTGCTT-GCCGAAGCCATGCTGGATAAAGAG	57
AY509004	tnpA_IS4	Mobility	Transposase A of insertion sequence IS*4*	AAGTTTGATTCCTTGGACTCTTCAGAATACA-GACAGCAAATAAAGACCTTTCGTTTGA	57
AY509004	tnpA_IS406	Mobility	Transposase A of insertion sequence IS*406*	AAATGTATGTCAAAGGAGTAAGTACCCGCAG-GGTCTCGGATATCGTCGAAATTCTTT	57
AF261825	tnpA_IS6100	Mobility	Transposase A of IS*6100*, SGI 1	CCGATCACGGAAAGCTCAAGATACTGAT-CAAGCCGGTGCGCGGTTTCAAATCGATCC	57
AF071413	tnpA_Tn21	Mobility	Transposase A of Tn*21*	GACTCCAAGGACGACCTGATCCGACATTACA-CATTCAACGATACCGACCTCTCGATC	58
AY509004	tnpA_Tn3	Mobility	Transposase A of Tn*3*	GTGTTCTTCAACCGCCTTGGGGAAATCAG-GGATCGGAGCTTCGAG	57
AP005147	tnpR_IS10	Mobility	Resolvase of insertion sequence IS*10*	CATGGTATAAATCCGTTGAGAAGCTGGGTTGG-TACTGGTTAAGTCGAGTAAGAGGAA	45
AF261825	tnpR_SG1	Mobility	SGI 1 resolvase	AGGGAGATTAGGGCATTACTCAAAGATGGTTC-TATTCCTGTATCTGATGTTGCTAGGCGA	57
AY144490	tpase1	Mobility	SPI 3, transposase 1 similar to transposase A	CGATTGTTAGGTTAAGGACACACCCTAATAC-CCCATTTGTTTCTGCTATCCTCAAAC	57

AE00647	traT	Mobility	pSLT plasmid- encoded conjugative transfer surface exclusion	GCGATGAGCACAGCAATCAAAAAGCGTAATCTT-GAAGTGAAAACCCAGATGAGTCAG	57
AF261825	trhH	Pathogenicity	SGI 1- encoded putative pilus assembly protein	ATAACAGCCTGCTTGAAGCCATGATGTC-TATAACTGGTGCCGTTGTCATTGGTGATT	57
M84642	wbaA_C1	Serotyping	O-antigen- polymerase: Serogroup C_1	GTGCTTGGTGCCATTCTATCATTGCCTTTGTCA-CATTATTTATCAGATATAATTTCTTCCGTT	63
X60665	wbaO_E1	Serotyping	Manosyl transferase (β 1-4 linkage): Sserogroups E_1 and D_2	ATCTTTGGATTTTAGGCTTGGCTCCCAGACCTT-TACTGAAGTTGATCGGCAAGTGTATATC	61
X56793	wbaU_B	Serotyping	Manosyl transferase (α 1-4 linkage): Serogroups B and D_1	TGCCTGATGCAATTTCCCGATTTAACAACTAT-GTCGCACGGTATGACTTTGATAATATGAAGC	63
X56793	wbaV_B	Serotyping	Abequosyl transferase: Serogroup B	CGGGTGTGATTTAGTTGAGATTAGAAACCCT-CATCGTTCTTGGCTCAGAACAGATGATGAAC	62
M65054	wbaV_D1	Serotyping	Tyvelosyl transferase: Serogroup D	TCGGCGATGGTTAAATGGTGGCAGTAGAT-TATTTTCTTTTAGCAATGAAGCTGATTTGA-TAGA	63
D14156	wcdA	Pathogenicity	UDP-glucose/GDP mannose dehy-drogenase	GATTATTGGGCTGGGATATGTTGGGCTTCCTCT-GGCAGTTGAATTTGGCAAATCTCG	57
D14156	wcdE	Pathogenicity	Required for translocation of the Vi polysaccharide to the cell surface	TACTCAAACAAGAGGATTGGGAGGGGCTAT-GCCTCTATTTCAGTCAGCATCCCGAAAA	58
D14156	wzf	Pathogenicity	Vi polysaccharide export inner-membrane protein	GATTCTGTCCGTAGAGCGTCATTAGTTATTAAA-GAGCGACGTGTTCAGCAAGCCAAG	59
AY334017	wzx_O6,14	Serotyping	O antigen flippase: O6,14 serovar factor	AAAGCGACCTTGAGTATTGGGCTCACTGCTG-TAGTAGTTATAATTTATAGTAGAGTGGA	57
AF017148	wzy(O27)	Serotyping	α- 1-6 polymerase: O27 serovar factor	TGAGTCTTTATTTAATCAAATATCTTTTATGCG-GATGCTGGATTGGCTACATCAAGGGGCAGT	63
AE008758	wzy_B	Serotyping	α- 1-2 polymerase: Serogroup B	TGGCGAATTACTCGGATTATACCCGTAATGCT-GTTCTTGTTGCTTCCTCAAACTTTG	57

(continued)

Table 1 (continued)

Accession No.	Probe name	Marker group	Gene function	Probe sequence	Length
U04165	wzy_D2/E1	Serotyping	Putative O antigen polymerase: Sserogroups D2 and Serogroup E1	AGGCGGTCAGTTTATATTCACAGAGGCTTTTCATTCACTTGGTTATGTCGGAGTATTCCTG	61
AE008706	yafD	Metabolism	Putative cytoplasmic protein	TGGTAGTAAATGTTCATGCGGTAAATTTTAGTCTGGGCGTGGACGTATACAGTAAGC	57
			Control oligonucleotides		
AF282268	ttrC	positive. Ccontrol	Tetrathionate reductase subunit C	ATGACGCATTCACTCATCATTGAAGAAGTGCTGGCTCACCCGCAGGACATTAGCTGG	57
X58149	PRKase	Negative Ccontrol	PRKase of *Arabidopsis thaliana*, negative control probe	TAACCTCTTCTTCTTCTTCCAAACAAGTCTTCCTCTACCGTCGTCAACCACAAACCA	57
M86720	rca	Negative Ccontrol	RCA of *Arabidopsis thaliana*, negative control probe	GATGATGAAGTGAGGAAGTTCGTTGAGAGCCTTGGAGTTGAGAAGATCGGAAAGAGG	57
NM_121758	rcp1	Negative Ccontrol	RCP1 of *Arabidopsis thaliana*, negative control probe	AGGTGTTAGGTTTGTAGGGTCTTTATCTGGATGGACAGCAACTCTTATGTTCATGTGGATG	57

Note: The table indicates the accession numbers, the probe names, corresponding marker group, a short description of the gene function, the probe sequence, and the oligonucleotide length of each probe. Sequence of SEN4287 is available from www.sanger.ac.uk/projects

References

1. Humphrey, T. (2000). Public-health aspects of *Salmonella* infection, in *Salmonella in Domestic Animals* (Way, C., and Way, A., eds.), CABI, Oxon, UK, pp. 245–263.
2. Selander, R. K., Beltran, P., Smith, N. H., Helmuth, R., Rubin, F. A., Kopecko, D. J., et al. (1990). Evolutionary genetic relationships of clones of *Salmonella* serovars that cause human typhoid and other enteric fevers. *Infect. Immun.* **58**, 2262–2275.
3. Wollin, R. (2007). A study of invasiveness of different *Salmonella* serovars based on analysis of the Enter-net database. *Euro Surveill.* **12**, 5–8.
4. Anonymous. (2007). Enter-net annual report: 2005—surveillance of enteric pathogens in Europe and beyond. Enter-net surveillance hub, HPA, Centre for Infections, Colindale, London.
5. Porwollik, S., and McCleeland, M. (2003). Lateral gene transfer in *Salmonella*. *Microbes Infect.* **5**, 977–989.
6. Levings, R. S., Lightfood, D., Partridge, S. R., Hall, R. M., and Djordjevic, S. P. (2005). The genomic island SGI1, containing the multiple antibiotic resistance region of *Salmonella enterica* serovar Typhimurium DT104 or variants of it, is widely distributed in other *Salmonella enterica* serovars. *J. Bacteriol.* **187**, 4401–4409.

Chapter 20

Methods for Data Analysis

William Paul Hanage and David Michael Aanensen

Abstract

The molecular epidemiology of infectious diseases uses a variety of techniques to assay the relatedness of disease-causing organisms to identify strains responsible for outbreaks or associated with particular phenotypes of interest (such as antibiotic resistance) and, it is hoped, provide insights into where and how these strains have emerged. The correct analysis of such data requires that we understand how the assayed variation accumulates. We discuss this with specific reference to three classes of methods: those based on gel electrophoresis of fragments generated by restriction enzymes or polymerase chain reaction (PCR), those based on microsatellites and other repeat elements, and raw sequence data from protein-coding genes. We also provide a simple example of how the likely origin of an apparently novel antibiotic-resistant strain may be identified and conclude with a discussion of some popular analysis packages and the more interesting prospects for the future in this rapidly developing field.

Key words: Analysis software, clustering, eBURST, homoplasy, molecular epidemiology, molecular typing, phylogenetics, population biology.

1. Introduction

The precise way in which molecular data are analyzed will obviously depend on the technique used. There are, however, some general features and considerations that we briefly discuss, along with some potential pitfalls.

Here are some typical questions in molecular epidemiology:

Is a case of disease due to an outbreak strain or an unrelated infection?

Is a case of tuberculosis, in a patient previously thought to be cured of the disease, due to reinfection with a new strain or due to a failure to clear the original infection?

D.A. Caugant (ed.), *Molecular Epidemiology of Microorganisms*, Methods in Molecular Biology, Vol. 551
DOI: 10.1007/978-1-60327-999-4_20,

How is an antibiotic-resistant strain of a pathogen related to those previously known? Is it a result of *de novo* acquisition of resistance?

Can we identify lineages associated with traits such as virulence?

2. The Importance of Clustering

The key to all of these questions is the relatedness of the isolates in the data set under study. Having used some methods (amply covered in the previous chapters) to assay genetic variation in the population of a pathogen, the next step is to cluster the organisms by relatedness. Analysis of molecular epidemiology data is mainly concerned with this. Depending on the type of variation assayed and the nature of the pathogen under study, there are several options available. One should be particularly aware of evolutionary processes that may make two isolates appear more related than they really are, such as recombination or convergent evolution. These processes will both lead to homoplasy—character states that are similar for reasons other than descent from a common ancestor.

To begin, consider two extreme outcomes of an investigation. At one end, all isolates are indistinguishable by the method used, while at the other end all are different and, moreover, equally different from each other. If either of these results is obtained, it is likely that the technique applied is not appropriate, offering in one case insufficient discrimination and in the other too much. While this may seem like a mere thought experiment, examples very like these two hypothetical data sets are to be found in the literature ***(1,2)***.

In a more successful study, some isolates will be indistinguishable, some will be very similar to them, and others will be very different. The processes by which we cluster these isolates require that we understand the way in which the assayed variation is generated so we can translate it into a genetic distance. Three types of variation are discussed next: (i) banding patterns produced by methods such as restriction fragment length polymorphisms (RFLPs) and amplified fragment length polymorphisms (AFLPs); (ii) microsatellites and other repeat elements; and (ii) raw DNA/RNA sequences.

2.1. Banding Patterns

Techniques such as pulsed-field gel electrophoresis (PFGE) and AFLP assay the presence of unique recognition sequences (sites where restriction enzymes cut or where polymerase chain reaction [PCR] primers bind) in the genome of a microorganism. The distances between these recognition sequences vary among strains, leading to DNA fragments of varying size produced by restriction digests or PCR with appropriate primers. These DNA fragments may be readily compared by electrophoresis, and the resulting banding patterns are the final data. Inherent to all these techniques is the ability to "bar code" the genomes of isolates of interest and the subsequent comparison of the bar codes provides the distance measure utilized for clustering.

The variation in this case is in the length of the DNA separating the recognition sequences. These sequences may be lost or gained by mutation, or indeed by recombination with related strains, and by other events, such as insertions, deletions, or large-scale genomic rearrangements that can alter the distances between recognition sequences. All of these mechanisms will alter the banding pattern.

As a result, these techniques offer very little information about the deep relatedness between strains with no recent common ancestor; unless two banding patterns are very similar, it is impossible to know how many events separate them or what those events might be. Moreover, the genes in which the assayed sequences lie are usually unknown; consequently some of the observed changes may be more likely to occur than others as a result of selection. For purposes such as outbreak analysis, this is not generally a problem as one is only attempting to identify the outbreak strain and any arising from it over the period of the outbreak. Therefore, one needs to assay variation that accumulates rapidly enough so that these can be distinguished from other unrelated infections, and consequently, one is looking for identical or very similar banding patterns. Interpretive criteria for how similar banding patterns should be for strains to be considered part of the same cluster vary between methods, but generally the banding patterns are required to show very high similarity. For an example, see the work of Tenover et al. ***(3)***.

By comparing the number of bands in common between profiles (inherently a somewhat subjective process that is in part automated in programs such as BioNumerics), it is possible to produce a distance matrix and from this a dendrogram (*see*, for example, ***ref. 4***). This should be avoided because it suggests more confidence about deep relationships than is appropriate given the considerations discussed.

2.2. Repeat Sequences

For organisms with limited sequence variation, including many eukaryotes, the existence of repeat elements in the genome offers an attractive source of variation for a typing method. In particular, the rates with which changes in these repeats occur are typically several orders of magnitude greater than mutations in protein-coding genes. As a result, organisms that are otherwise difficult to type because they are almost identical may be readily resolved. An example in bacteria is spoligotyping of *Mycobacterium tuberculosis* ***(5)*** and in several fungal species multilocus microsatellite typing (MLMT) ***(6)***. When analyzing data from such methods, it is important as ever to consider the mechanisms by which variation arises. For example, variation at a single microsatellite is constrained to a minimum and maximum number of repeats. Given the high rate of mutation, it is not at all unlikely that at a single locus, an identical allele (i.e., same number of repeats) could arise

by chance. This will produce homoplasy (*see* **Subheading 2.**) and has an effect similar to recombination in that it can make two distantly related isolates appear more closely related than they actually are. Adding more loci may improve the resolution, and this is the rationale underlying the inclusion of multiple microsatellites in MLMT. The limits of the size of the microsatellite array also mean that genetic distances at microsatellites become saturated more quickly than variation in coding sequences. For example, estimates of the divergence time between the two fungal species *Coccidioides immitis* and *Coccidioides posadasii* obtained from microsatellites, on one hand, and their flanking regions, on the other hand, differed by an order of magnitude (7.6×10^5 and 12.8×10^6 yr, respectively). As expected, this discrepancy was not found when considering more closely related populations *(7)*.

In the case of other target sequences, such as the variable direct repeats assayed by spoligotyping, it is thought that variation arises through deletion. Therefore, theoretically, one should be able to compare closely related sequences and not only cluster them but also infer the direction of change (because those with fewer direct repeats must be derived from those with more). However, in this and similar cases, phylogenies cannot be created because the repeats violate the assumption of independence among the variable sites: Adjoining repeats may be lost together *(8)*. A final general comment about repeat regions of this kind and others is that their function is rarely understood. Hence, the selective consequences of variation at these loci are an open question.

2.3. Sequence Data

With the increasing availability and decreasing expense of DNA sequencing, the direct determination of sequence data is becoming more and more prominent as a method for epidemiological typing. It might be thought that this makes data analysis easier than the examples given, but this is not necessarily so.

It is now known that many bacteria undergo frequent horizontal transfer of homologous genes *(9)*. As a result, if we examine a single locus and find two isolates that are identical at that locus, this may be no reflection of the overall relatedness of these isolates because the relevant genetic material could have been imported relatively recently. This (and other reasons) has led to the development of multilocus sequence typing (MLST) *(10)*, by which sequences are determined for multiple loci. This buffers against the distorting effect of horizontal gene transfer: Even if one locus changes through recombination, then the others do not and remain to give a better account of the relationship between the isolate in question and the rest of the population.

If the recombination rate is high enough, it is again very difficult to draw conclusions about deep branches in the tree. For many species, a dendrogram produced from the pairwise differences between MLST profiles contains minimal phylogenetic information beyond relatively close linkage distances. While it is possible

to produce a tree from concatenates of the sequences at the MLST loci, this has the problem that a single recombinational import can introduce many nucleotide changes. As a result, in addition to problems with the branching order or topology of the tree *(11)*, branch lengths can be artificially inflated as many polymorphisms are introduced by a single event. Two approaches to this problem are the programs eBURST *(12)* and CLONALFRAME *(13)*. The former focuses on changes at the very tips of the tree, and the latter attempts to identify and account for recombinational imports.

eBURST is discussed in detail elsewhere *(14–16)* and in the example given in **Subheading 3.3.** It handles recombination by focusing on single-allele changes and weighting them equally whether they introduce a single base change or many (the latter is likely to be recombination). In contrast, CLONALFRAME attempts to identify recombination events and account for them by simply excluding them from the analysis. The sequence which remains is the 'clonal frame'. In simulations it performs well, but it makes the assumption that recombination imports arise from *outside* the sample. Its performance in a situation when most recombinant alleles are present in the sample is not clear. Both programs become less reliable with very high recombination rates *(13,16)*.

No investigator should overlook recombination when interpreting data and deciding on a method of analysis. However, in some species, such as *Staphylococcus aureus*, it appears to be sufficiently rare that conventional phylogenies can be constructed with few qualms (although, as usual, care should be taken to select and deploy an appropriate model of nucleotide substitution). Such organisms may also show considerable stability in terms of clonal structure. Thus, the same clones can be recognized over a long period of time, and band-based methods produce similar results to MLST *(17)*. In very rapidly mutating and highly recombinogenic organisms, such as *Helicobacter pylori*, this is not possible, and every epidemiologically unconnected strain is different.

Similar considerations apply to viruses for which recombination can also be frequent and confuse matters. RNA viruses require particular attention when reconstructing phylogenies. In contrast to many of the problems faced by phylogeneticists, in molecular epidemiology we usually consider closely related members of the same species. This means that for bacteria, for example, one can be more relaxed about problems that arise when considering very distantly related taxa (such as long-branch attraction; *18,19)*. RNA viruses, however, are the most rapidly evolving organisms known: Substantial viral diversity arises in a single human immunodeficiency virus (HIV) 1 patient. Properly accounting for different rates of substitution and selection is of great importance in this field and is far beyond the scope of this chapter. A useful start point for interested readers is **ref.** *20,* and investing in a phylogenetics textbook (e.g., **ref.** *19*) is recommended if such techniques form a large part of future plans.

3. Example: Origin of Resistant Clones of *Streptococcus pneumoniae*

3.1. Background

Streptococcus pneumoniae (the pneumococcus) is a major childhood pathogen. A vaccine has recently been devised that is highly effective at preventing disease due to 7 of its more than 90 serotypes (4, 6B, 9V, 14, 18C, 19F, 23F). Vaccination has almost totally removed the vaccine serotypes from the population, including many important antibiotic-resistant clones, which expressed vaccine serotypes *(21)*.

Pneumococcal clones can change their serotypes by recombination events that insert the relevant serotype loci from other clones, a process known as *serotype switching (22)*. As a result, simply recording serotype is insufficient to tell us how a given isolate is related to others. While vaccination initially has a marked and beneficial effect on the prevalence of antibiotic resistance (because vaccine serotypes were more likely to be associated with resistance) *(23)*, investigators are beginning to record increased resistance to antimicrobials among serotypes not included in the vaccine.

In 2007, Pichichero et al. reported a pneumococcal isolate retrieved from a case of ear infection that was resistant to all classes of antibiotics licensed for pediatric use in the United States. The serotype of the isolate was 19A, a nonvaccine serotype *(24)*. This finding raised the following question: Has resistance been acquired *de novo,* or does this clone arise from serotype switching allowing a vaccine-serotype-resistant clone to acquire a new, nonvaccine serotype (in this case 19A)?

3.2. Choice of Method and Results

Techniques such as PFGE are unlikely to enable us to say what this strain has derived from because, in the pneumococcus, variation assayed by this method accumulates very quickly. Instead, MLST, which assays sequence variation at multiple sites around the genome and assigns an allelic profile (*see* **Chapter 11**), is the method of choice. Variation at these sites accumulates relatively slowly, so it should be informative regarding the relationship over the longer term. The MLST database also contains a large number of strains with which the results can be compared.

Subjecting the isolate in question to MLST, it was found to have the allelic profile 7, 11, 10, 15, 6, 8, 1, corresponding to sequence type (ST) 2722. To find whether there are any other records of this ST, we can go to the MLST database at http://spneumoniae.mlst.net/ and enter this ST via the advanced query page (http://spneumoniae.mlst.net/advanced/). Doing so reveals no additional records from strains with that ST isolated prior to the time of writing (January 15, 2008). This means that this is the first time an ST 2722 strain has been reported. If there was another strain in the database with a vaccine serotype and a similar resistance profile, it could have suggested that ST 2722 had been derived from it by serotype switching.

However, to conclude with certainty that serotype switching has not been involved, we must consider how ST 2722 is related to the rest of the pneumococcal population and consider whether there are any other likely parent strains.

3.3. Clustering: eBURST

eBURST groups genotypes defined by MLST or similar methods into "clonal complexes" through a simple parsimony-based method. Briefly, clonal complexes are defined on the basis of genetic similarity: All members of a clonal complex must share genetic information to a specified degree of similarity (which may be specified by the user) with at least one other member. Within clonal complexes, the genotype with the largest number of minor variants is assigned as the most likely ancestor, and the patterns of descent from this ancestor are defined according to specified rules.

To find whether eBURST can predict a putative ancestor for ST 2722, simply visit http://spneumoniae.mlst.net/eburst/ and select the option to run eBURST on the whole *S. pneumoniae* MLST database. Once the program has opened, select the analysis window and click "compute" using the default parameters. The group or clonal complex into which ST 2722 falls is then simply located using the "Find ST" feature illustrated in **Fig. 1**. At the time of writing, ST 2722 falls within group 1 (*see* **Subheading 3.5.**).

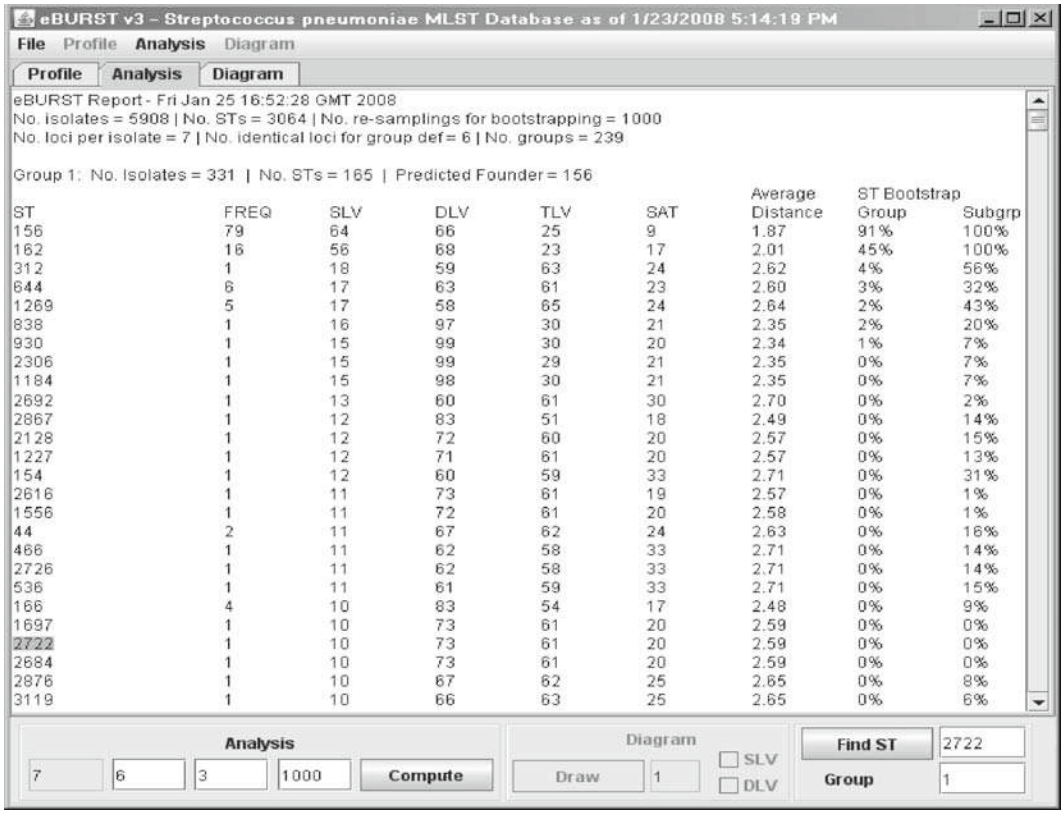


Fig. 1. Textual output from eBURST analysis of the entire pneumococcal MLST database. The "Find ST" function is at bottom right of the window.

3.4. Graphical Representation

To produce a graphical representation of the group containing ST 2722, select the Diagram window, enter the group number in the dialogue box at the bottom of the page, and click draw. The results should resemble **Fig. 2**. Again, ST 2722 may be located using the Find ST feature as shown.

The blue circle in the center represents ST 156, and those arranged around it are the STs differing from ST 156 at one of the seven MLST loci. The nature of the difference (single-base change or many) is not taken into consideration. Hence, this method is refractory to distortion by recombination.

eBURST implies that the most likely ancestor of this clonal complex, and therefore of ST 2722, is ST 156. This is interesting because this is the ST of one of the major multiresistant clones circulating prior to introduction of the vaccine *(25)*. Specifically, it is the Spain 9V-3 clone. As implied in its name, the serotype of the majority of the strains with this ST is 9V, one of the targeted vaccine serotypes.

By clicking on ST 156 and selecting "database" from the diagram pull-down menu as shown, we may link to the MLST

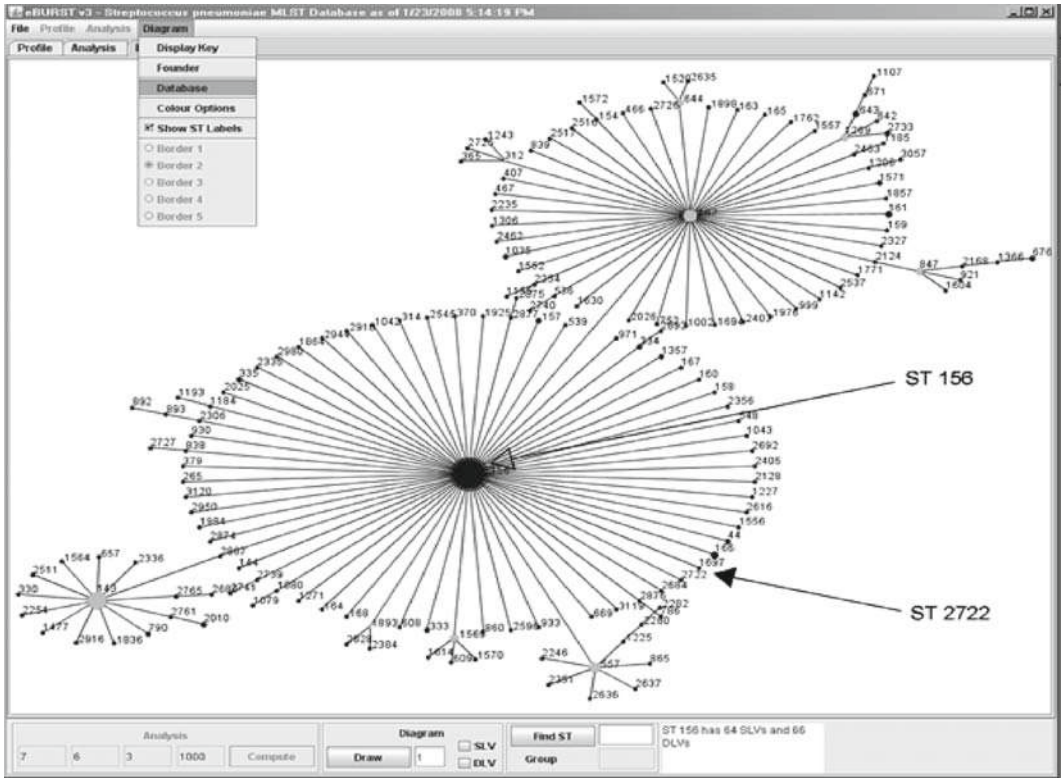


Fig. 2. Screenshot showing Group 1, the largest clonal complex in the pneumococcal MLST database at the time of writing. The predicted overall ancestor of the clonal complex ST 156 and its single-locus variant ST 2722 are both indicated. The pull-down menu at top left shows how the MLST database may be accessed within eBURST.

database. The results retrieved are those that have been submitted to the MLST database by sources worldwide. As can be seen, to date these do not include 19A strains. We can therefore suggest that the multiresistant phenotype of ST 2722 was inherited from its ancestor, ST 156, but that it has acquired a new capsule suited to the postvaccine era.

3.5. Caveats

While it is almost certainly true that ST 2722 is derived from an ST 156 strain, there are several caveats that should be mentioned. At present no isolates of ST 2722 prior to that reported by Pichichero et al. *(24)* have been recorded, but they may be in the future and then provide a more accurate picture of how, where, and when this lineage acquired a new capsule. At the time of writing, ST 2722 fell within the largest clonal complex in the MLST database. As more records are entered, the composition of this group will change, and in time it may surrender its position as the largest clonal complex. It should also be appreciated that considerable variation could be present within ST 156 (all we know for sure is that all ST 156 strains have identical sequences at seven gene fragments).

Finally, the clonal complex in question illustrates how sampling can bias analyses. Because ST 156 is highly resistant, it has been frequently reported to the MLST database, along with numerous minor variants, like ST 2722, which are also resistant. This inflates the number of resistant strains in the database and the number of minor variants of ST 156. As a result, eBURST identifies ST 156 as the ancestor of this clonal complex. However, the true overall ancestor, which gave rise to ST 156, is ST 162. This ST is not, however, a multiresistant clone, so it and its susceptible variants are underrepresented in the database. For a more thorough discussion of this issue, *see* **ref. *15.***

4. Looking Forward

A survey of this kind is inevitably destined to be obsolete in a few years as new analysis methods and opportunities arise and become widespread. Here, we summarize some of the methods that seem particularly interesting.

4.1. Population Analysis

The genetic data that are collected for molecular epidemiology are suitable, in fact ideal, for approaches that aim to identify discrete populations within a species. From the point of view of epidemiology, and when combined with data on the infected hosts, this can conceivably define groups within which contact is more likely. The proof of principle for this is a study of *H. pylori*

in which a close concordance was found between the strains colonizing human hosts and the ethnic origin of those human hosts *(26)*. The reason for this relation is that *H. pylori* is normally acquired from only very close contacts, often vertically, and hence can be used to recapitulate human population movements. Popular programs for this sort of analysis include STRUCTURE *(27)* and BAPS *(28,29)*. An excellent review of these and others, together with a discussion of what a population identified by such methods actually is, may be found in **ref. 30**.

4.2. Spatial Analysis

For many infections, one may identify certain strains or lineages of the pathogen that are endemic to a particular region. In some cases, this may be the result of clear phylogeographic structure, while in other cases it may be due to seeding of an epidemic from elsewhere. In both circumstances, it is helpful to be able to visualize the spatial location of the cases and the associated genotypes of the disease-causing organisms. This is becoming possible through several initiatives that link epidemiological data with mapping and other data (e.g., GENELAND; *31*). One possible limitation of such analyses is concern over linking disease data to a specific location (e.g., methicillin-resistant *S. aureus* within a particular hospital), which may limit their applications; compromises will have to be negotiated.

4.3. Inferring Evolutionary History from Trees

Trees can tell us more than how closely taxa are related. They can also be used to estimate the time at which two lineages diverged, given assumptions about the rate with which substitutions accumulate. Moreover, the distribution of mutations on the tree (which relate to the lengths of the branches in question) will vary depending on the selective and demographic history of the population. What can we learn from examining the shape of the genealogy?

As a first step, some statistical description of the shape of the genealogy is needed. This is offered by the coalescent. This approach, developed by Kingman *(32)*, describes the distribution of "coalescent events" going back in time from the present. Coalescent events are where two lineages coalesce, for instance, internal nodes in a genealogy. These represent the point at which two lineages shared a common ancestor. The way the rate of coalescence relates to variation in the population size is illustrated in **Fig. 3**. In the case of sequences from a population that experienced a severe bottleneck (**Fig. 3a**), the part of the genealogy that corresponds to this time will have a higher number of coalescent events because the chance of two lineages sharing a common ancestor becomes higher in the smaller population during the bottleneck. In the case of a rapidly expanding population, the tree will resemble **Fig. 3b**, with long terminal branches and short internal ones.

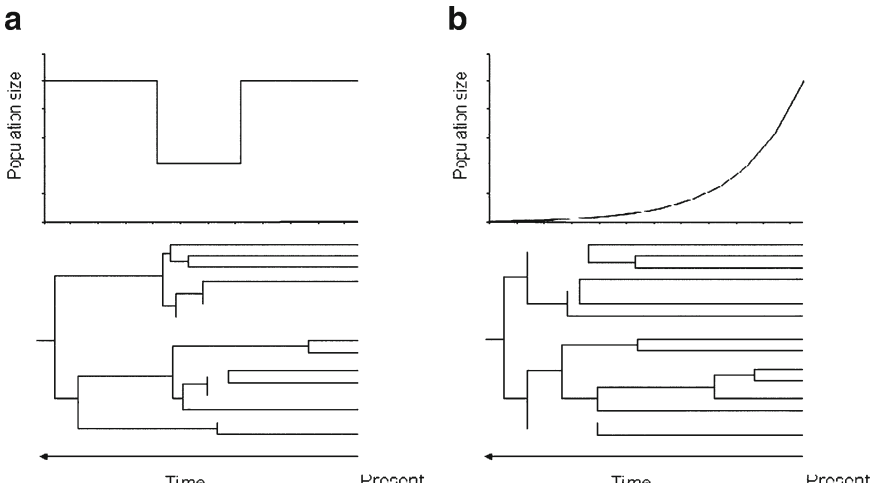


Fig. 3. Two examples of how changing population size can affect the shape of the genealogy of a set of sequences. Time is depicted on the *x* axis. (**a**) The consequences of a population bottleneck. The majority of the coalescent events take place in the part of the genealogy corresponding to the period of the bottleneck, shown here directly below it. (**b**) A population growing in exponential fashion. Again, the smaller population is associated with an increased rate of coalescence, leading to a genealogy in which the majority of terminal branches are quite long.

A fertile field in modern epidemiology is using these insights to explore the history of a set of sequences and to produce "skyline plots" *(33)*, which track the population size over the history of the tree (for some nice examples of this approach, *see* **refs.** *34* and *35*). In a particularly exciting development, it has become possible to incorporate information about the time at which a sample was collected and to study what are sometimes called "measurably evolving populations," such as HIV-1 infecting an individual host *(36)*. A comprehensive discussion of coalescent theory is far beyond the scope of this chapter, but interested readers are referred to one of the several books (e.g., *37*,*38*) becoming available on this exciting topic.

4.4. Population Genomics and Single-Nucleotide Polymorphisms

It is becoming ever easier to sequence a genome, especially if at least one genome of the species in question has already been sequenced. While whole-genome comparisons, as a matter of routine for epidemiology, are still some way off, these studies do furnish us with a comprehensive catalogue of the sites at which the reference genomes differ. Single-nucleotide polymorphisms (SNPs) so identified may be targeted and used as the basis of a typing scheme. When considering SNP data one should, as ever, consider the evolutionary forces generating the assayed differences (is there evidence of homoplasy or selection, for instance? Are the assayed SNPs at synonymous sites?). A further, very important consideration in SNP analysis is phylogenetic discovery

bias. Under the approach outlined, our typing scheme will seek out only those SNPs that differ between the sequenced strains, that is, those that occurred on the evolutionary path separating the reference strains. Just as a microarray based on one or a few genomes cannot, by definition, be used to detect or study genes that are not present in those genomes, this approach cannot identify diversity that is present within branches leading to unsequenced strains. This problem is less acute for organisms in which recombination is common because here the genomic variation present in the rest of the species will have been sampled along the evolutionary path separating the reference strains. Of course, recombination leads to other problems, as discussed in **Subheading 2.3.** Discovery bias has been demonstrated in both real and theoretical situations *(39, 40)* and may be combated by a polymorphism discovery approach in which many genes are sampled in diverse isolates (for an example, *see* **ref.** *41*).

4.5. Work Flows and Web Services

Programmatic interfaces for many common bioinformatics procedures, such as BLAST and CLUSTAL, are located on Web servers across the globe (e.g., the National Center for Biotechnology Information [NCBI] has Web services for accessing database searches under the EFetch facility). This allows software developers to build tools that utilize these procedures without the need to program such functions themselves. Data are simply sent to the service provider and the results returned in an appropriate format. This method of reusable functionality is similar in concept to the approach utilized at the repository of modules for the freely distributed R statistics software found at CRAN (http://cran.r-project.org/). However, rather than downloading and using modules on a single computer, Web services allow the development of truly distributed software. A Web service can be described at its most basic as a set of three parts: (i) a definition of inputs (e.g., a FASTA file containing a set of sequences); (ii) the algorithms utilized on the inputs (e.g., a CLUSTAL alignment); and (iii) the outputs produced from step ii (e.g., an alignment file and tree definition file).

To allow less programmatically inclined users to access such facilities, programs have been developed that provide graphical interfaces to the wealth of Web services available. One of the more promising examples is Taverna *(42)*, developed by the European Bioinformatics Institute (EBI). Taverna allows multiple Web services to be chained together into a work flow, using a graphical interface to define a series of steps that are undertaken on an initial input set of data. A work flow is built up as a flowchart with the inputs for each step defined and the outputs passed into the next step of the work flow. The resulting work flow is viewed as a flowchart, and the work flow "definition" can be saved as a text file, allowing reuse.

More usefully, the work flow can be uploaded to a Web site that allows others to take advantage, test, and amend the original user's experimental steps (see http://www.myexperiment.org). Many standard procedures involved in molecular epidemiological studies could be developed in such a fashion, with the advantage of user amendments, review, and testing. An example of such a work flow can be seen at http://www.myexperiment.org/workflows/124. This allows a single novel sequence to be entered along with a list of EMBL accession numbers for sequences on which a homology search is to be performed. Sequences are retrieved from GenBANK, translated into protein sequences, followed by an all-versus-all BLAST to identify homology. Subsequently, a CLUSTAL alignment is undertaken on a nonredundant set of BLAST matches and, finally, a neighbor-joining (NJ) tree or unweighted pair group method with arithmetic means (UPGMA) tree is produced along with the alignment files as final output. Should anyone wish to run this work flow, they can simply paste a URL into Taverna, and the work flow is imported. Access to the hundreds of Web services available means work flows are only limited by the users' questions. Combined with peer review of submitted work flows, this offers a very powerful complement to current software development methodologies.

5. Packages for Analysis of Molecular Epidemiological Data

A multitude of options exists when it comes to programs for data analysis. Some are summarized in **Table 1**, which presents a list that is by no means exhaustive; it focuses on those mentioned in this chapter. Most of the commonly utilized software packages have overlapping functionality, and very similar analyses can be undertaken using more than one package. As mentioned, interpretation, particularly when inferring relationships between strains and population-level analysis, is dependent on understanding the algorithms used. Some of the criticisms directed at published results are more likely due to inappropriate use of methodology rather than at the programs themselves.

One distinctive feature of many molecular epidemiology laboratories is high throughput of samples, with a requirement that experiments be conducted and results recorded in a standardized fashion. Some packages contain additional features to help in project management. These are generally the proprietorial packages, which may be quite expensive. However, many free alternatives are available for the majority of analyses. The advantages of the former, in addition to the inclusion of laboratory information management software (LIMS), is clear

Table 1
Programs Mentioned in the Text

Name	Web site	Functions	Cost	Comments
BioNumerics	http://www.applied-maths.com/bionumerics/bionumerics.htm	LIMS and numerous bioinformatics modules including MST and 1 and 2D gel analysis	$$$	Includes scripting language
CLCBio	http://www.clcbio.com/	LIMS-MLST module available for additional fee	$$$	
Geneious	http://www.clcbio.com/	LIMS	$$	Includes scripting language
Phineus	http://www.phineus.org/	LIMS and MLST analysis		In development
eBURST	http://eburst.mlst.net/	Patterns of recent descent		
CLONALFRAME	http://www2.warwick.ac.uk/fac/sci/statistics/staff/research/didelot/clonalframe/	Accounting for recombination in genealogies		
BAPS	http://web.abo.fi/fak/mnf//mate/jc/software/baps.html	Identifying populations and admixture between them using genetic data		
STRUCTURE	http://pritch.bsd.uchicago.edu/structure.html	Identifying populations and admixture between them using genetic data		
MEGA	http://www.megasoftware.net/	Phylogenetics		
PAUP*	http://paup.csit.fsu.edu/	Phylogenetics	$	
PAML	http://abacus.gene.ucl.ac.uk/software/paml.html	Phylogenetics		
MrBayes	http://mrbayes.csit.fsu.edu/	Phylogenetics		
PHYLIP	http://evolution.genetics.washington.edu/phylip.html	Phylogenetics		
BEAST	http://beast.bio.ed.ac.uk/			
SPLITSTREE	http://www.splitstree.org	Production of reticulate phylogenies that display conflict in the data		Includes Neighbor-net

Note: This is not an exhaustive list, and new options are continually becoming available.

documentation and user support, which in the case of free packages is inevitably limited by how much time the developer has available to devote to it.

In terms of proprietorial packages, the industry leader is BioNumerics. A tool for bioinformatics applications in general, BioNumerics has several modules specifically tailored to epidemiological investigations (including MLST and *spa* typing). SQL databases are used to store results from a remarkable variety of tests, including sequences, 1D and 2D gels, metabolic tests, microarray data, and more. Furthermore, BioNumerics includes a scripting language that allows users to manipulate core features of the software programmatically. While undoubtedly powerful, most laboratories will probably struggle to make use of many of its features. Alternative packages include CLCBio, which is again focused on general bioinformatics applications. However, plug-ins for epidemiological applications are becoming available (e.g., MLST).

Of stand-alone phylogenetics programs, the best known is probably MEGA *(43)*, with PAML *(44)*, MrBayes *(45,46)*, PAUP* *(47)*, and PHYLIP *(48)* also popular for certain specific purposes and within some user communities. The most recent version of MEGA *(43)*, still free to download, contains an integrated trace file editor and alignment tools. It uses a relatively intuitive interface with dialogue boxes and pull-down menus. In contrast, most other packages are operated via the command line, which can nevertheless be useful (e.g., if preparing batch analysis procedures, obviating the need for repeated clicking). A common feature of all these is that while they can take sequences or distance matrices as input, results from gel-based methods must be converted to a distance matrix beforehand. The BEAST *(49)* package (with related programs) is a well-supported, easy-to-use program for the sorts of analyses briefly discussed in **Subheading 4.3.** A recent and excellent review has surveyed the population genetics packages available together with their strengths and weaknesses *(50)*. While not all of these will be useful to researchers in this field, those who are interested in population genetic questions will find them invaluable.

In representing sequences where recombination is known or suspected, methods such as CLONALFRAME *(13)* or eBURST *(12)* should be used, but the drawbacks of each of these should be appreciated. CLONALFRAME, in detecting anomalous DNA, must assume that it arises from outside the sample, which may produce problems when both donor and recipient are present in the sample. eBURST, in contrast, tells you nothing about the relatively deep branches that separate the clonal complexes identified. Both programs perform poorly under

conditions of very high recombination. BioNumerics allows the construction of minimum-spanning trees (MSTs), which build on eBURST results to link clonal complexes via hypothetical unsampled intermediates. How secure this assumption is and whether it introduces an additional source of error into such analyses remain to be tested. Finally, programs such as Splitstree or Neighbor-net *(51,52)* find a reticulate phylogeny for the data and so represent recombinant genealogies better than a bifurcating tree. However, it should be noted that the signal that these programs detect may be produced by processes other than recombination, and that they are on their own not an acceptable test for recombination.

6. Concluding Remarks

The analysis of data is as important as their collection, and it is important that the correct tools are used and their underlying assumptions made plain. The discussion in this chapter focuses on clustering because this is the most important aspect in the majority of epidemiological questions. The degree to which microorganisms can be clustered and the appropriate tools to do it depend on the type of variation assayed, the rate of recombination, and so on. It is increasingly common for data collected as part of epidemiological investigations to be used for population genetic purposes, that is, to investigate hypotheses regarding the evolution of pathogens. In the opposite direction, insights derived from population genetics may be used to study pathogen spread.

Whichever technique is applied and however the data are analyzed, one should remember that a program of laboratory work should not be undertaken without a thorough understanding of the procedures that are to be used. This applies just as much to the tools that are used to interpret the data.

Acknowledgments

W. P. H. gratefully acknowledges the support of the Royal Society. D. M. A. is funded by a Wellcome Trust program grant awarded to Brian Spratt. We would like to thank Mat Fisher for helpful discussions.

References

1. Go, M. F., Kapur, V., Graham, D. Y., and Musser, J. M. (1996). Population genetic analysis of *Helicobacter pylori* by multilocus enzyme electrophoresis: extensive allelic diversity and recombinational population structure. *J. Bacteriol.* **178**, 3934–3938.
2. Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A., and Carniel E. (1999). *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proc. Natl. Acad. Sci. U S A* **96**, 14043–14048.
3. Tenover, F. C., Arbeit, R. D., and Goering, R.V. (1997). How to select and interpret molecular strain typing methods for epidemiological studies of bacterial infections: a review for healthcare epidemiologists. *Infect. Control Hosp. Epidemiol.* **18**, 426–439.
4. Singh, A., Goering, R. V., Simjee, S., Foley, S. L., and Zervos, M. J. (2006). Application of molecular techniques to the study of hospital infection. *Clin. Microbiol. Rev.* **19**, 512–530.
5. Aranaz, A., Liebana, E., Mateos, A., Dominguez, L., Vidal, D., Domingo, M., et al. (1996). Spacer oligonucleotide typing of *Mycobacterium bovis* strains from cattle and other animals: a tool for studying epidemiology of tuberculosis. *J. Clin. Microbiol.* **34**, 2734–2740.
6. Fisher, M. C., Aanensen, D., de Hoog, S., and Vanittanakom, N. (2004). Multilocus microsatellite typing system for *Penicillium marneffei* reveals spatially structured populations. *J. Clin. Microbiol.* **42**, 5065–5069.
7. Fisher, M. C., Koenig, G., White, T. J., and Taylor, J. W. (2000). A test for concordance between the multilocus genealogies of genes and microsatellites in the pathogenic fungus *Coccidioides immitis*. *Mol. Biol. Evol.* **17**, 1164–1174.
8. Warren, R. M., Streicher, E. M., Sampson, S. L., van der Spuy, G. D., Richardson, M., Nguyen, D., et al. (2002). Microevolution of the direct repeat region of *Mycobacterium tuberculosis*: implications for interpretation of spoligotyping data. *J. Clin. Microbiol.* **40**, 4457–4465.
9. Thomas, C. M., and Nielsen K. M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* **3**, 711–721.
10. Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U S A* **95**, 3140–3145.
11. Feil, E. J., Holmes, E. C., Bessen, D. E., Chan, M. S., Day, N. P., Enright, M. C., et al. (2001). Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc. Natl. Acad. Sci. U S A* **98**, 182–187.
12. Feil, E. J., Li, B. C., Aanensen, D. M., Hanage, W. P., and Spratt, B. G. (2004). eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J. Bacteriol.* **186**, 1518–1530.
13. Didelot, X., and Falush, D. (2007). Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**, 1251–1266.
14. Feil, E. J., and Enright, M. C. (2004). Analyses of clonality and the evolution of bacterial pathogens. *Curr. Opin. Microbiol.* **7**, 308–313.
15. Spratt, B. G., Hanage, W. P., Li, B., Aanensen, D. M., and Feil, E. J. (2004). Displaying the relatedness among isolates of bacterial species—the eBURST approach. *FEMS Microbiol. Lett.* **241**, 129–134.
16. Turner, K. M., Hanage, W. P., Fraser, C., Connor, T. R., and Spratt, B. G. (2007). Assessing the reliability of eBURST using simulated populations with known ancestry. *BMC Microbiol.* **7**, 30.
17. Melles, D. C., van Leeuwen, W. B., Snijders, S. V., Horst-Kreft, D., Peeters, J. K., Verbrugh, H. A., et al. (2007). Comparison of multilocus sequence typing (MLST), pulsed-field gel electrophoresis (PFGE), and amplified fragment length polymorphism (AFLP) for genetic typing of *Staphylococcus aureus*. *J. Microbiol. Methods* **69**, 371–375.
18. Felsenstein, J. (1978). Cases in which parsimony of compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401–410.
19. Felsenstein, J. (2004). *Inferring Phylogenies* (Sunderland, M. A., ed.). Sinauer Associates, Sunderland, MA.
20. Hall, B. G., and Barlow, M. (2006). Phylogenetic analysis as a tool in molecular epidemiology of infectious diseases. *Ann. Epidemiol.* **16**, 157–169.
21. Hanage, W. P. (2007). Serotype replacement in invasive pneumococcal disease: where do we go from here. *J. Infect. Dis.* **196**, 1282–1284.
22. Coffey, T. J., Daniels, M., Enright, M. C., and Spratt, B. G. (1999). Serotype 14 variants of the Spanish penicillin-resistant serotype 9V clone of *Streptococcus pneumoniae* arose by large recombinational replacements of the cpsA-pbp1a region. *Microbiology* **145**, 2023–2031.
23. Kyaw, M. H., Lynfield, R., Schaffner, W., Craig, A. S., Hadler, J., Reingold, A., et al. (2006). Effect of introduction of the pneumococcal conjugate vaccine on drug-resistant

Streptococcus pneumoniae. *N. Engl. J. Med.* **354**, 1455–1463.

24. Pichichero, M. E., and Casey, J. R. (2007). Emergence of a multiresistant serotype pneumococcal strain not included in the 7-valent conjugate vaccine as an otopathogen in children. *JAMA* **298**, 1772–1778.
25. Zhou, J., Enright, M. C., and Spratt, B. G. (2000). Identification of the major Spanish clones of penicillin-resistant pneumococci via the Internet using multilocus sequence typing. *J. Clin. Microbiol.* **38**, 977–986.
26. Falush, D., Wirth, T., Linz, B., Pritchard, J. K., Stephens, M., Kidd, M., et al. (2003). Traces of human migrations in *Helicobacter pylori* populations. *Science* **299**, 1582–1585.
27. Falush, D., Stephens, M., and Pritchard J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587.
28. Corander, J., and Tang, J. (2007). Bayesian analysis of population structure based on linked molecular information. *Math. Biosci.* **205**, 19–31.
29. Tang, J., Tao, J., Urakawa, H., and Corander, J. (2007). T-BAPS: a Bayesian statistical tool for comparison of microbial communities using terminal-restriction fragment length polymorphism (T-RFLP) data. *Stat. Appl. Genet. Mol. Biol.* **6,** Article 30.
30. Waples, R. S., and Gaggiotti, O. (2006). What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Mol. Ecol.* **15**, 1419–1439.
31. Guillot, G., Estoup, A., Mortier, F., and Cosson, J. F. (2005). A spatial statistical model for landscape genetics. *Genetics* **170**, 1261–1280.
32. Kingman, J. F. C. (1982). On the genealogy of large populations. *J. Appl. Probability* **19A**, 27–43.
33. Pybus, O. G., Rambaut, A., and Harvey P. H. (2000). An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* **155**, 1429–1437.
34. Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22**, 1185–1192.
35. Salemi, M., de Oliveira, T., Soares, M. A., Pybus, O., Dumans, A. T., Vandamme, A. M., et al. (2005). Different epidemic potentials of the HIV-1B and C subtypes. *J. Mol. Evol.* **60**, 598–605.
36. Ewing, G., Nicholls, G., and Rodrigo, A. (2004). Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in measurably evolving populations. *Genetics* **168**, 2407–2420.
37. Hein, J., Schierup, M. H., and Wiuf, C. (2005). *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, Oxford, UK.
38. Nielsen, R. (2004). Statistical methods in molecular evolution, in *Statistics for Biology & Health*, Springer-Verlag, New York.
39. Pearson, T., Busch, J. D., Ravel, J., Read, T. D., Rhoton, S. D., U'Ren, J. M., et al. (2004). Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing. *Proc. Natl. Acad. Sci. U S A* **101**, 13536–13541.
40. Keim, P., Van Ert, M. N., Pearson, T., Vogler, A. J., Huynh, L. Y., and Wagner, D. M. (2004). Anthrax molecular epidemiology and forensics: using the appropriate marker for different evolutionary scales. *Infect. Genet. Evol.* **4**, 205–213.
41. Roumagnac, P., Weill, F. X., Dolecek, C., Baker, S., Brisse, S., Chinh, N. T., et al. (2006). Evolutionary history of *Salmonella typhi*. *Science* **314**, 1301–1304.
42. Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M. R., Li, P., et al. (2006). Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.* **34**, W729–W732.
43. Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599.
44. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.
45. Ronquist, F., and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574.
46. Huelsenbeck, J. P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755.
47. Swofford, D. L. (2003). *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Sinauer Associates, Sunderland, MA.
48. Felsenstein, J. (1989). PHYLIP—phylogeny inference package (Version 3.2). *Cladistics* **5**, 164–166.
49. Drummond, A. J., and Rambaut A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214.
50. Excoffier, L., and Heckel, G. (2006). Computer programs for population genetics data analysis: a survival guide. *Nat. Rev. Genet.* **7**, 745–758.
51. Huson, D. H. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68–73.
52. Bryant, D., and Moulton V. (2004). Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.* **21**, 255–265.

Chapter 21

Internet-Based Sequence-Typing Databases for Bacterial Molecular Epidemiology

Keith A. Jolley

Abstract

As the use of nucleotide sequence-based typing has become more widespread in the investigation of microbial epidemiology, there has been a natural requirement for curated Internet-based databases that can act as central authorities for nomenclature and type definitions. These facilitate the sharing and comparison of data between laboratories without the need for reference samples. Here, the use of the most common multilocus sequence typing (MLST) and antigen sequence databases are described. In particular, for MLST, the steps required for allele sequence and profile identification are explained along with a detailed overview of searching and matching isolate records. BLAST searching of antigen sequence databases is also described.

Key words: Bacterial typing, databases, Internet, MLST, nucleotide sequencing.

1. Introduction

The genotyping of strains is a central requisite of global molecular epidemiology and is required for the detailed study of transmission dynamics. In an outbreak situation, strain identification of the etiological agent is useful for determining treatment regimens and prophylactic measures, but sometimes the overriding requirement is to determine whether two isolates are either identical or clonally related to each other. For this, many highly discriminatory comparative techniques with continuous values may be used, such as pulsed-field gel electrophoresis (PFGE), restriction fragment length polymorphism, or random amplification of polymorphic DNA. Continuous values, such as the measurement of

D.A. Caugant (ed.), *Molecular Epidemiology of Microorganisms*, Methods in Molecular Biology, Vol. 551
DOI: 10.1007/978-1-60327-999-4_21,

electrophoretic mobility of proteins or DNA in a gel, suffer from potential ambiguity resulting from the level of precision that the measurement can be made. Measurements can be affected by extraneous factors, so their use requires standard markers to calibrate the experiment if comparisons are to be made with existing data. This is problematic for long-term epidemiology and portability of data, particularly the use of online databases.

In contrast, epidemiological surveillance requires that the same biological markers are used routinely so that results are reproducible over time and between laboratories. Such library typing techniques preferably make use of markers with categorized values, for example, DNA sequences in multi- or single-gene sequence typing or integers representing the number of repeats in variable-number tandem repeat analysis. Databases that use continuous values of data do exist for molecular epidemiology; PulseNet *(1)*, utilizing PFGE data for enteric bacteria, is perhaps the best-known example, although its use is computationally intensive, and the uploading of large image files for processing is required to make comparisons.

Increasingly, the use of nucleotide sequence technology is becoming predominant in surveillance as it offers significant advantages over competing methodologies, namely, in reproducibility, portability of data, and resolution. Nucleotide sequences are highly amenable to electronic transmission and storage in databases, while a range of freely available software is available to facilitate its comparison. With near-ubiquitous access to the Internet, the use of online sequence databases for rapid identification and comparison of microorganisms has increased rapidly.

A distinction should be made between archival and actively curated databases. Archival databases, such as GenBank, often accept direct submissions without oversight of data quality and store data for publication and general identification purposes. These are invaluable for identifying genes and species but have little to offer for epidemiological purposes. Actively curated databases that have been set up specifically for strain typing, however, such as those for multilocus sequence typing (MLST) or for specific antigen genes, are essential for accurate microbial identification required for surveillance.

2. Methods

2.1. Multilocus Sequence Typing

MLST indexes the neutral variation in sequences of housekeeping gene fragments *(2,3*; *see also* **Chapter 11** for practical details). The use of multiple loci provides a robust method of typing organisms that undergo frequent recombination that

would otherwise invalidate phylogenetic methods using a single locus. As with any method that relies on nucleotide sequence data, results are unambiguous and portable, making them amenable to electronic storage. MLST databases are now available for at least 40 organisms, mainly bacteria, ensuring a uniform nomenclature. More than half of these are hosted at the University of Oxford in the United Kingdom (http://pubmlst.org) ***(4)***, with other schemes hosted at the United Kingdom's Imperial College (http://www.mlst.net) ***(5)***; the Environmental Research Institute, Cork, Ireland (http://mlst.ucc.ie); and the Pasteur Institute, Paris (http://www.pasteur.fr/mlst/).

There are three main types of query that MLST databases address: (i) allele sequence identification and comparison; (ii) allelic profile or sequence type (ST) identification and comparison; and (iii) matching of isolates. All MLST Web databases offer these functions, but the exact steps required and additional functions available vary. Here, the steps used on the PubMLST and Pasteur sites that use the mlstdbNet software ***(4)*** and the mlst.net site are specifically described.

2.1.1. Allele Sequence Identification

After trace files have been assembled, they are generally trimmed so that the sequence starts and finishes at the endpoints of the defined MLST locus under consideration. This can be done manually (*see* **Note 1**) or by using automated tools such as STARS (http://sara.molbiol.ox.ac.uk/userweb/mchan/stars/) or Phineus (http://www.phineus.org). Identification of the allele can then be determined as described next.

2.1.1.1. PubMLST Site

1. Select "Single locus query" within the profiles database.
2. Choose the appropriate locus, paste the sequence into the Web form, and click the "Submit Query" button.
3. If the sequence has been defined previously, the identity of the matching allele will be displayed along with a "Find similar" link to discover similar alleles and the nucleotides at which they vary. If the sequence has not been defined, the Web site will display the name of the allele to which it is most similar, along with a list of the nucleotides that vary so they can be confirmed (*see also* **Note 2**).

2.1.1.2. mlst.net Site

1. Select "Single Locus" from the "Locus Query" drop-down list box on the database front page.
2. Choose the appropriate locus, paste the sequence into the Web form, and click the "Submit" button.
3. If the sequence has been defined previously, the identity of the matching allele will be displayed. If the sequence is new, a message will inform you of the most similar allele along with its percentage identity.

4. To determine the nucleotide differences between the query sequence and the nearest match, click the "Sequence analysis" button. This opens a Java applet window with the query sequence aligned with known alleles. Visual inspection of this alignment will identify where the sequences vary (*see* **Note 3**).
5. Confirmed new sequences should be submitted to the database curator for inclusion.

2.1.2. Allelic Profile Identification and Comparison

Once alleles have been identified for each of the loci, the ST can be determined.

2.1.2.1. PubMLST Site

The batch profile function of the profiles database provides the easiest method for profile determination (even if you are only determining the ST for one profile).

1. Select "Batch profile query" within the profile database.
2. Copy and paste the sample identifier and allelic profile directly from a spreadsheet into the Web form. You can copy as many samples as you wish together; each sample should be on a separate row with columns separated by any amount of white space.
3. Click "Submit." A table displaying the sample identifier, allelic profile, ST, and clonal complex, if appropriate, will be returned (*see* **Note 4**).

2.1.2.2. mlst.net Site

1. Select "Allelic" from the "Profile Query" drop-down list box on the database front page.
2. Enter the allele numbers for each locus of your profile in the appropriate boxes of the form.
3. Ensure that the query type is set for "Exact or nearest match."
4. Click "Query list of distinct STs." A table will be displayed showing either an exact match, if available, or the nearest matching profiles otherwise.

2.1.3. Searching the Isolate Databases

The MLST isolate databases offers various search capabilities to find isolates that match any criteria of interest.

2.1.3.1. PubMLST Site

1. Click "Search database" from the isolate database front page.
2. The search form allows values to be selected from a number of fields that can be combined so that either all or any (and/or) are matched. The values can be specified to match exactly, match partially, to be greater or less than, to be not, or to not contain the selected value (*see* **Note 5**). Results can be ordered by any field, and the number of records per page can be set. To search for an empty field, the value "<blank>" can be specified. Click "Submit" once search criteria have been entered (*see* **Note 6**).
3. A page of results is returned. To navigate to the next set of results, click the ">" button on the page bar at the bottom of

the page. Further information about any particular isolate can be found by clicking its hyperlinked ID number.

4. The results can be broken down by individual fields by clicking the "Breakdown dataset" button. This generates charts for each field showing the frequency for each field value. More detailed analyses can be performed by clicking the "Advanced breakdown" button. This provides options to break the data set down against any two fields, so, for example, the frequency of strain type fields can be determined by country. The frequency of field combinations can also be shown where any selection of fields can be chosen. This can be particularly useful, for instance, in determining the surface antigen repertoire of a collection of isolates.

2.1.3.2. mlst.net Site

1. Select "Database query" from the "Profile Query" drop-down list box on the database front page.
2. The search form allows values to be selected from a number of fields that can be combined so that either all or any are matched. The values can be specified to match exactly (not case sensitive), to be greater or less than, or to be not. Click "Submit" (*see* **Note 7**).
3. The query results will be displayed in a table (*see* **Note 8**). More detailed information about each isolate can be found by clicking the hyperlinked ID number.

2.2. Antigen Sequence Typing

Sequencing of antigen genes is being used increasingly in place of serological characterization of isolates in bacterial typing schemes. Nucleotide sequencing has the advantage that every variant can be identified, whereas many isolates can be nontypable using monoclonal antibody panels. Curated Web databases for specific antigens, such as the *Neisseria meningitidis* serotyping (PorB) ***(6)***, serosubtyping (PorA) ***(7)*** and FetA ***(8)*** proteins, *Campylobacter jejuni* FlaA ***(9)*** and MOMP, *Streptococcus zooepidemicus* seM ***(10)***, and *Wolbachia* Wsp ***(11)*** proteins are available. These databases make use of the agdbNet software ***(12)***, which offers BLAST ***(13)*** querying of nucleotide or peptide variants and linking to isolate data.

The sequence query page of these databases allows either a nucleotide or a peptide sequence to be entered and compared to all known alleles or variants using the BLAST algorithm. This works whether the variants are defined by their nucleotide or peptide sequence. In some databases, multiple loci may be defined, and these can all be queried at the same time, or the user can specifically select the locus to search against. Depending on the database, isolates with a matching antigen variant can also be retrieved.

1. Select "Single sequence query" (or in some databases "Identify variable region").

2. Select the locus of interest if the database contains multiple loci. Alternatively, a value of "all" will search all loci together.
3. Copy and paste either a nucleotide or a peptide sequence into the Web form and click "Submit Query."
4. If identical matches are found, these will be listed (*see* **Note 9**) along with a link to the BLAST results output. If no exact matches are found, a list of partial matches will be displayed along with their percentage identity, the number of mismatches, the number of gaps, and the length of the alignment.
5. Results are hyperlinked, so clicking these will navigate to further information about the particular sequence or variant. This may include GenBank accession numbers, publications in which the sequence is described, or links to matching isolates.

3. Notes

1. The "locus explorer—polymorphic site analysis" function of a PubMLST profiles database shows a schematic of a particular MLST locus, clearly identifying start and end points and showing all known mutations within the gene fragment, colored by their relative abundance within defined alleles. This information can be very useful when trimming allele sequences manually.
2. On the PubMLST site, MLST sequences can also be queried against all known alleles using BLAST, available within the profiles database. This has the advantage that sequence trimming is not required, and it is not necessary to specify the locus. This is important if the start point cannot be identified and confirmation is needed that the sequence obtained from the sequencer is the correct locus and not something else due to a mix-up of samples. The disadvantage is that one does not get a simple list of nucleotide differences to the nearest known allele.
3. In three of the databases on www.mlst.net (*Enterococcus faecalis*, *Staphylococcus epidermidis*, *Streptococcus pneumoniae*), the single-locus query offers choices of "Simple results" or "View DNA mismatches." The first simply states the nearest allele and its percentage identity in the case of an unknown allele sequence. The latter displays a list of the defined alleles along with the positions of nucleotides that vary. If available, the latter display makes confirming nucleotide differences easier than using the "Sequence analysis" applet.
4. Using the batch profile function of a PubMLST database to identify STs from profiles has a major advantage over the standard "allelic profile query" in that allele numbers do not need to be

manually transcribed into the Web form. The standard allelic profile query does, however, allow partial matching of profiles to be performed to identify related STs.

5. In a PubMLST isolate database, if combinations of four fields are not sufficient, the interface can be customized by going to the options page and changing the number of fields to include (up to 20 fields can be used). If a field is of particular interest, the interface can be customized further to include a drop-down box of that field's values to be selected. These options are remembered between sessions, so the interface can be customized to a user's preferences.
6. Isolate data sets can also be retrieved from a PubMLST database by using a list query. Here a long list of attributes, usually ID numbers or isolate names, can be pasted into a Web form. This makes it particularly easy to retrieve specific records, especially if a list has already been prepared.
7. In a database on mlst.net there is no documented way to search for an empty field.
8. On larger databases on mlst.net, exercise restraint in making queries that are likely to return a large proportion of the database. The system does not offer a way to break the results into multiple pages, and you may find your Web browser locks up when attempting to display hundreds or thousands of rows of data.
9. BLAST searches of antigen databases can sometimes return multiple exact matches. This occurs if variants have been defined that are identical to preexisting sequences with the exception of missing end motifs, sometimes seen in variants defined by a variable-loop sequence. The software uses the BLAST algorithm to simply identify the presence of an exact sequence and does not have a definition of the start or end points of the variant or allele. In cases such as these, the longest matched sequence is usually the correct one, but the user should check by trimming the sequence back to the defined end points of the allele and then query again. The correct variant will be an exact match and the same length as the query sequence.

References

1. Swaminathan, B., Barrett, T. J., Hunter, S. B., and Tauxe, R. V. (2001). PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerg. Infect. Dis.* **7**, 382–389.
2. Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U S A* **95**, 3140–3145.
3. Maiden, M. C. (2006). Multilocus sequence typing of bacteria. *Annu. Rev. Microbiol.* **60**, 561–588.
4. Jolley, K. A., Chan, M. S., and Maiden, M. C. (2004). mlstdbNet—distributed multi-locus

sequence typing (MLST) databases. *BMC Bioinformatics* **5**, 86.

5. Aanensen, D. M., and Spratt, B. G. (2005). The multilocus sequence typing network: mlst.net. *Nucleic Acids Res.* **33**, W728–W733.
6. Urwin, R., Russell, J. E., Thompson, E. A., Holmes, E. C., Feavers, I. M., and Maiden, M. C. (2004). Distribution of surface protein variants among hyperinvasive meningococci: implications for vaccine design. *Infect. Immun.* **72**, 5955–5962.
7. Russell, J. E., Jolley, K. A., Feavers, I. M., Maiden, M. C., and Suker, J. S. (2004). PorA variable regions of *Neisseria meningitidis*. *Emerg. Infect. Dis.* **10**, 674–678.
8. Thompson, E. A. L., Feavers, I. M., and Maiden, M. C. J. (2003). Antigenic diversity of meningococcal enterobactin receptor FetA, a vaccine component. *Microbiology* **149**, 1849–1858.
9. Dingle, K. E., Colles, F. M., Ure, R., Wagenaar, J., Duim, B., Bolton, F. J., et al. (2002). Molecular characterisation of *Campylobacter jejuni* clones: a rational basis for epidemiological investigations. *Emerg. Infect. Dis.* **8**, 949–955.
10. Kelly, C., Bugg, M., Robinson, C., Mitchell, Z., Davis-Poynter, N., Newton, J. R., et al. (2006). Sequence variation of the SeM gene of *Streptococcus equi* allows discrimination of the source of strangles outbreaks. *J. Clin. Microbiol.* **44**, 480–486.
11. Baldo, L., Dunning Hotopp, J. C., Jolley, K. A., Bordenstein, S. R., Biber, S. A., Choudhury, R. R., et al. (2006). Multilocus sequence typing system for the endosymbiont *Wolbachia pipientis*. *Appl. Environ. Microbiol.* **72**, 7098–7110.
12. Jolley, K. A., and Maiden, M. C. (2006). AgdbNet—antigen sequence database software for bacterial typing. *BMC Bioinformatics* **7**, 314.
13. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.

INDEX

D

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z