Population Genomics

Martin F. Polz
Om P. Rajora *Editors*

Population Genomics: Microorganisms

Springer

Population Genomics

Editor-in-Chief
Om P. Rajora
Faculty of Forestry and Environmental Management
University of New Brunswick
Fredericton, NB, Canada

This pioneering *Population Genomics Series* deals with the concepts and approaches of population genomics and their applications in addressing fundamental and applied topics in a wide variety of organisms. Population genomics is a fast emerging discipline, which has created a paradigm shift in many fields of life and medical sciences, including population biology, ecology, evolution, conservation, agriculture, horticulture, forestry, fisheries, human health and medicine.

Population genomics has revolutionized various disciplines of biology including population, evolutionary, ecological and conservation genetics, plant and animal breeding, human health, genetic medicine, and pharmacology by allowing to address novel and long-standing intractable questions with unprecedented power and accuracy. It employs large-scale or genome-wide genetic information across individuals and populations and bioinformatics, and provides a comprehensive genome-wide perspective and new insights that were not possible before.

Population genomics has provided novel conceptual approaches, and is tremendously advancing our understanding the roles of evolutionary processes, such as mutation, genetic drift, gene flow, and natural selection, in shaping up genetic variation at individual loci and across the genome and populations, disentangling the locus-specific effects from the genome-wide effects, detecting and localizing the functional genomic elements, improving the assessment of population genetic parameters or processes such as adaptive evolution, effective population size, gene flow, admixture, inbreeding and outbreeding depression, demography and biogeography, and resolving evolutionary histories and phylogenetic relationships of extant and extinct species. Population genomics research is also providing key insights into the genomic basis of fitness, local adaptation, ecological and climate acclimation and adaptation, speciation, complex ecologically and economically important traits, and disease and insect resistance in plants, animals and/or humans. In fact, population genomics research has enabled the identification of genes and genetic variants associated with many disease conditions in humans, and it is facilitating genetic medicine and pharmacology. Furthermore, application of population genomics concepts and approaches can facilitate plant and animal breeding, forensics, delineation of conservation genetic units, understanding evolutionary and genetic impacts of resource management practices and climate and environmental change, and conservation and sustainable management of plant and animal genetic resources.

The volume editors in this Series have been carefully selected and topics written by leading scholars from around the world.

Martin F. Polz • Om P. Rajora
Editors

Population Genomics: Microorganisms

Springer

Editors
Martin F. Polz
Department of Civil and Environmental Engineering
Massachusetts Institute of Technology
Cambridge, MA, USA

Om P. Rajora
Faculty of Forestry and Environmental Management
University of New Brunswick
Fredericton, NB, Canada

ISSN 2364-6764 ISSN 2364-6772 (electronic)
Population Genomics
ISBN 978-3-030-04755-9 ISBN 978-3-030-04756-6 (eBook)
https://doi.org/10.1007/978-3-030-04756-6

Library of Congress Control Number: 2018965734

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Dedicated to my mentors from whom I have learnt, and the student and postdoctoral colleagues from whom I continue to learn.

Martin F. Polz

Respectfully dedicated to my educators and mentors.

Om P. Rajora

Preface

Genomics has revolutionized many fields of biology. For microbes, in particular, it has revealed the enormous scope of diversity coexisting in most environments. Not surprisingly, efforts in microbial genomics have, to a large extent, been directed towards understanding the phylogenetic and functional diversity encompassed by microbes. Although much of microbial diversity remains to be uncovered, there is also a more recent focus on analysis of closely related genomes. This effort was, at least initially, driven by the need to better understand the evolution and epidemiology of pathogenic viruses and bacteria. The continuous decline in sequencing cost has, however, enabled a broader focus on nonhuman pathogens, and environmental and industrial microbes to better understand how microevolutionary processes create variation within populations. Hence, the field of microbial population genomics has come of age, and we believe, it is time for a book that summarizes current developments and future perspectives in this novel but important field.

Population genomics of microorganisms is most commonly understood to encompass the analysis of entire genomes of intraspecific and interspecific closely related individuals using phylogenetic and population genetics concepts and tools. Population genomics, therefore, deals broadly with the analysis of evolutionary forces that both create and remove variation among members of populations, and, perhaps most importantly, lead to adaptation to environmental niches or hosts. Simply put, population genomics is population genetics empowered by genomics. This definition can, however, vary somewhat according to the organisms studied so that many authors within this book provide their own, more nuanced definitions of microbial population genomics. Moreover, availability of data varies greatly for different types of organisms. Not surprisingly, the genomic analysis of human viral and bacterial pathogens is most advanced and although other fields are catching up, for many types of organisms, population genomics of microorganisms represents a nascent field emerging from comparative genomics of closely related organisms. The availability of large collections of closely related strains is, however, bound to rapidly increase in the next few years since standard genetic characterization of

isolates is increasingly done by whole genome rather than single-marker gene sequencing.

In this book, we have tried to cover all major groups of microorganisms for which at least some population genomics studies have been undertaken. The chapters, thus, span the whole spectrum of diversity encompassed by microbes, including bacteria, archaea, fungi, and viruses. And for pathogens, there is further subdivision according to the hosts infected. The result is an in-depth analysis of microbial population genomics that allows comparison among fields. Our hope is that this structure will enable the reader to find commonalities and differences among organisms, and that such comparison will outline a roadmap for new investigators in the field of microorganism population genomics. Because crosstalk between fields is always mediated by common methods, we have included a chapter that explicitly deals with computational tools for microorganism population genomics. However, many of the individual chapters cover additional methods, often developed for specific purposes but often more broadly relevant. Finally, because many microbes remain hard to culture and are only accessible by metagenomics, the book contains a chapter that deals explicitly with the opportunities and challenges in applying population genomics to uncultured organisms.

Talking about population genomics implies that we know how to define and delineate populations. In many cases, we have good intuition of what a population might be, such as in the analysis of highly clonal pathogens or sexually isolated eukaryotes. How to demarcate population boundaries is, however, often not easy. In particular, for bacteria and archaea, as well as for some viruses, the potential for horizontal gene transfer and the vast coexisting genetic diversity exemplify this difficulty. In fact, the term population is often used loosely in microbiology, describing from cells in a culture tube to diverse microbes coexisting in environmental samples. Several of the chapters, therefore, explicitly tackle the issue of how to define populations and how populations split into distinct units in the process of speciation. Based on the sophistication of the analysis, we predict that the next few years will see tremendous advances in theory about how to define microbial populations.

It is an exciting time for a book on microbial population genomics as the field takes shape and is expanding into new areas of research. We thank all the distinguished authors who have taken the time to contribute to this effort and we hope that all have been rewarded by the timeliness and quality of this book.

Cambridge, MA, USA — Martin F. Polz
Fredericton, NB, Canada — Om P. Rajora

Contents

Contributors

Philip Arevalo Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

David A. Baltrus School of Plant Sciences, University of Arizona, Tucson, AZ, USA

School of Animal and Comparative Biomedical Sciences, University of Arizona, Tucson, AZ, USA

Bridget Marie Barker The Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, USA

Harald Brüssow Division of Animal and Human Health Engineering, Laboratory of Gene Technology, University of Leuven, Leuven, Belgium

Guillaume Charron Département de Biologie and Département de Biochimie, Microbiologie et Bio-informatique, Institut de Biologie Intégrative et des Systémes, PROTEO, Université Laval, Quebec City, QC, Canada

Marcus de Melo Teixeira The Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, USA

Vincent J. Denef Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA

Kanny Diallo Department of Zoology, University of Oxford, Oxford, UK

Centre pour les Vaccins en Développement, Bamako, Mali

Xavier Didelot Department of Infectious Disease Epidemiology, Imperial College London, London, UK

Éléonore Durand Département de Biologie and Département de Biochimie, Microbiologie et Bio-informatique, Institut de Biologie Intégrative et des Systémes, PROTEO, Université Laval, Quebec City, QC, Canada

Chris Eberlein Département de Biologie and Département de Biochimie, Microbiologie et Bio-informatique, Institut de Biologie Intégrative et des Systémes, PROTEO, Université Laval, Quebec City, QC, Canada

Fernando García-Arenal Centro de Biotecnología y Genómica de Plantas UPM-INIA, Universidad Politécnica de Madrid, Madrid, Spain

E.T.S. Ingeniería Agronómica, Alimentaria y de Biosistemas, Madrid, Spain

Fernando González-Candelas Joint Research Unit "Infection and Public Health" FISABIO-Universitat de València, Institute for Integrative Systems Biology, I2SysBio (CSIC-UV), Valencia, Spain

CIBER in Epidemiology and Public Health, Madrid, Spain

Mathieu Hénault Département de Biologie and Département de Biochimie, Microbiologie et Bio-informatique, Institut de Biologie Intégrative et des Systémes, PROTEO, Université Laval, Quebec City, QC, Canada

Daniel Raymond Kollath The Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, USA

David J. Krause Laboratory of Genetics, Genome Center of Wisconsin, Wisconsin Energy Institute, J. F. Crow Institute for the Study of Evolution, University of Wisconsin-Madison, Madison, WI, USA

Christian R. Landry Département de Biologie and Département de Biochimie, Microbiologie et Bio-informatique, Institut de Biologie Intégrative et des Systémes, PROTEO, Université Laval, Quebec City, QC, Canada

Martin C. J. Maiden Department of Zoology, University of Oxford, Oxford, UK

Hélène Martin Département de Biologie and Département de Biochimie, Microbiologie et Bio-informatique, Institut de Biologie Intégrative et des Systémes, PROTEO, Université Laval, Quebec City, QC, Canada

Lou Nielly-Thibault Département de Biologie and Département de Biochimie, Microbiologie et Bio-informatique, Institut de Biologie Intégrative et des Systémes, PROTEO, Université Laval, Quebec City, QC, Canada

Israel Pagán Centro de Biotecnología y Genómica de Plantas UPM-INIA, Universidad Politécnica de Madrid, Madrid, Spain

E.T.S. Ingeniería Agronómica, Alimentaria y de Biosistemas, Madrid, Spain

Juan Ángel Patiño-Galindo Joint Research Unit "Infection and Public Health" FISABIO-Universitat de València, Institute for Integrative Systems Biology, I2SysBio (CSIC-UV), Valencia, Spain

CIBER in Epidemiology and Public Health, Madrid, Spain

Martin F. Polz Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

B. Jesse Shapiro Department of Biological Sciences, University of Montreal, Montreal, QC, Canada

Carlos Valiente-Mullor Joint Research Unit "Infection and Public Health" FISABIO-Universitat de València, Institute for Integrative Systems Biology, I2SysBio (CSIC-UV), Valencia, Spain

CIBER in Epidemiology and Public Health, Madrid, Spain

David VanInsberghe Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Rachel J. Whitaker Department of Microbiology, Carl R. Woese Institute for Genomic Biology, University of Illinois Urbana-Champaign, Champaign, IL, USA

Part I
Concepts and Approaches

Computational Methods in Microbial Population Genomics

Xavier Didelot

Abstract Whole genome sequencing is frequently applied to hundreds of samples within a single microbial population study. The resulting datasets are large and need to be analysed using computationally efficient methods, the development of which is an active research field. Here we review the current state of the art in terms of computation methods used in microbial population genomics. This includes software for assembly and alignment of core genomic regions, which is usually a pre-requirement for analysing the ancestry of the genomes, via phylogenetic on non-phylogenetic methods. We also review additional techniques aimed at combining genomic data with temporal, geographical or other types of metadata, as well as pan-genome methods of analysis that go beyond the core genome.

Keywords Alignment • Assembly • Computation methods • Microbial population genomics • Pan-genome analysis • Phylodynamics • Phylogenetics • Phylogeography • Recombination

1 Introduction

With the advent of new genome sequencing technologies, the cost and time required to sequence whole microbial genomes have decreased to such a point that research studies are now able to include hundreds or even thousands of newly sequenced genomes. Analysis of such large datasets requires the use of specific computational methods, which are reviewed in this chapter, but are still the subject of active development. Section 2 describes how to prepare genomic data for

X. Didelot (✉)
Department of Infectious Disease Epidemiology, Imperial College London, Norfolk Place, London W2 1PG, UK
e-mail: x.didelot@imperial.ac.uk

Martin F. Polz and Om P. Rajora (eds.), *Population Genomics: Microorganisms*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2017_3,

analysis, including identification of core and accessory genomic regions, assembly and alignment. Section 3 summarises methods for analysing the ancestry of the genomes, which can broadly be divided into phylogenetic and non-phylogenetic approaches. Section 4 describes how temporal information about the sampling dates of the genomes can be combined with the genomic data to paint a more complete picture of the evolutionary process. Section 5 covers the use of the geographic locations from which the genomes originate to describe the geographic structure of the population. Section 6 describes how other types of metadata can be integrated into a microbial population genomics study to investigate the distribution and evolution of various properties of interest. Finally, Sect. 7 explains how analysis of the pan-genome can be carried out.

2 Preparing Genomic Data for Analysis

2.1 Core and Accessory Genome

When comparing genomic data from members of a microbial population, it is useful to identify the genomic regions that are present in all the genomes, and which collectively are called the core genome. The remaining regions, which are found in some but not all the genomes, are collectively called the accessory genome, while the sum of core and accessory genome is often called the pan-genome. Analysis of microbial population genomic data typically requires an alignment of the core genomic regions, and this section describes how to prepare such an alignment. The separation of core and accessory genome regions is especially relevant for bacterial population genomics, because bacterial genomes within a population often exhibit significant variation in genomic content, whereas this is not usually the case for viral populations. In bacterial genomics, analysis of non-core regions can be important too, and this subject is treated in Sect. 7.

2.2 Reference-Based Assembly

Sequencing data from current sequencing instruments (reviewed in Loman and Pallen 2015; Goodwin et al. 2016) comes in the form of a large number of reads of length 100–250 bp which are highly redundant, so that each individual genomic position is covered by several reads. The average number of reads covering genomic positions is called the coverage depth and is a good indication of how accurate the final genome sequence will be, for example depth of 40× and above. Assembly is the process whereby reads are put together to reconstruct the genome sequence. There are broadly two forms of assembly: reference-based assembly and de novo assembly, each with their specific strengths and weaknesses.

Reference-based assembly requires that a whole genome from the population (or at least species) under study has been previously sequenced, which is called the reference genome. Each read is then aligned against the reference genome, and popular algorithms to perform this include BWA (Li and Durbin 2009), SMALT (http://www.sanger.ac.uk/science/tools/smalt-0), Stampy (Lunter and Goodson 2011) and Bowtie (Langmead et al. 2009). The next step is called variant calling, which is often done using SamTools (Li et al. 2009), FreeBayes (Garrison and Marth 2012) and/or GATK (McKenna et al. 2010). For each position along the reference genome, the alignment of reads at that position is considered. If there are enough aligned reads and they are in good agreement, the corresponding nucleotide of the target genome is called, otherwise it is left undetermined. The latter happens mostly for regions that are present in the reference genome but not in the target genome and for regions that are repetitive, or if the sequencing quality is low. Reference-based assembly has the advantage that each assembled genome is aligned against the same reference and hence all aligned against each other and directly comparable. Drawbacks include the need for a pre-sequenced reference genome, and the fact that only regions found in the reference genome can be assembled, which is sufficient to study the core genome but not the accessory genome.

2.3 De Novo Assembly

The alternative to reference-based assembly is to assemble each genome de novo, that is by directly comparing and aligning the reads against each other. Popular softwares for de novo assembly include Velvet (Zerbino and Birney 2008), SPAdes (Bankevich et al. 2012), IDBA (Peng et al. 2012) and A5 (Tritt et al. 2012). The output is typically a set of long genomic regions called contigs, which occur along the genome in an undetermined order. De novo assembled genomes need to be aligned against each other before they can be compared. A first approach is to perform a multiple alignment of the whole genome which accounts for possible genomic rearrangements, but even the best software using this strategy such as progressiveMauve (Darling et al. 2010) or MUGSY (Angiuoli and Salzberg 2010) cannot deal with more than ~50 genomes. An alternative is to align each de novo assembled genome against a single reference, for example using MUMmer (Kurtz et al. 2004), but this shares the disadvantages of reference-based assembly described above. A third approach to using de novo assembled data is to search for previously defined genes throughout the contigs using for example BLAST (Altschul et al. 1997), as implemented for example in the BIGSdb platform (Jolley and Maiden 2010). Finally, instead of using predefined genes it is possible to annotate each de novo assembled genome separately, using for example RAST (Overbeek et al. 2014), Prokka (Seemann 2014) or Prodigal (Hyatt et al. 2010) and to search for orthologs between the genomes using a pipeline involving BLAST to compare the genes versus each other, for example OrthoMCL (Li et al. 2003),

LS-BSR (Sahl et al. 2014) or Roary (Page et al. 2015). Once ortholog genes have been found in all genomes, they can be aligned separately using for example Muscle (Edgar 2004).

Reference-based and de novo assemblies are complementary approaches which are often used side by side to compare results in ambiguous regions and exploit the strengths of both strategies, especially the reconstruction of core-genome alignments that are directly analysable in reference-based assembly and the reconstruction of non-core regions in de novo assembly. After applying either or both approaches, an alignment of the genomes is created which contains all core regions (or core genes if a de novo gene-based approach was used). Such an alignment is required as input for the analytical methods described in the next sections.

2.4 Simulation

Analysis of simulated microbial genomic data in parallel with real genomic data can often be useful. This allows for example to test the fit of an evolutionary model to the data, to build empirical distributions of expected quantities, or to estimate evolutionary parameters informally by progressive tuning of the simulation parameters until it resembles the real data. A more formal use of simulated datasets is to use Approximate Bayesian Computation techniques, also known as likelihood-free methods since they do not require to calculate the probability of the data given evolutionary parameters, but instead rely on simulation and comparison of the simulated and real data on a set of summary statistics (Marin et al. 2012). Simulation is also useful on its own (i.e. without combination with real data), to test the accuracy of analytical methods on datasets for which the correct answer is known.

The most popular and powerful approaches to simulate microbial genomic data are based on the coalescent model (Kingman 1982) under which it is possible to simulate the evolutionary history of a sample of genomes without simulating the evolution of the whole population. One of the first methods to be implemented based on this principle was Hudson's ms (Hudson 2002), and it remains popular to date due to the wide range of scenarios that can be simulated using this software. Extensions have also been released, for example msHOT (Hellenthal and Stephens 2007) which allows for the presence of mutational hotspots. Another popular software is fastsimcoal (Excoffier and Foll 2011), which uses an efficient approximation to simulate crossover recombination, allowing the simulation of longer genomes from sexual populations. Clonal organisms such as bacteria undergo a recombination process akin to gene conversion rather than crossover, which can be simulated for example in ms but for which separate algorithms have been specifically implemented. SimMLST (Didelot et al. 2009b) was aimed at simulating multi-locus sequence typing data, where sequence is available for only a handful of short (~400 bp) gene fragments (Maiden et al. 1998). It has recently been superseded by SimBac (Brown et al. 2016) which is 100 times faster and therefore much better suited to simulate whole genome sequence data.

3 Description of Microbial Population Ancestry

3.1 Phylogenetics Ignoring Recombination

The most frequently used method to represent patterns of relationships between a set of microbial genomes is to draw a phylogenetic tree (Fig. 1). Closely related genomes should have fewer differences between them and be more closely clustered together on the tree compared to more distantly related genomes. A tree should always be read along the axis from root to leaves, bearing in mind that the other axis is arbitrary so that two genomes can be next to each other and yet be separated by a long branch (Baum et al. 2005). Phylogenetic methods are computational techniques that use as input an alignment of genomic data like the ones described in the previous section, and produce in output a phylogenetic tree. Most phylogenetic methods assume that no recombination happened, which is appropriate for example to analyse data from bacterial pathogens do not recombine much, e.g. *Mycobacterium tuberculosis* (Comas et al. 2013), or data in which recombinant regions have been previously detected and filtered out (cf next section).

The simplest phylogenetic methods rely on first building a distance matrix between all pairs of genomes, for example UPGMA, Neighbor-Joining (Saitou and Nei 1987) and BIONJ (Gascuel 1997). These methods are not very popular because they do not exploit the full data but only the distance matrix. They are however very quick and therefore still frequently used to provide a starting point for other methods. Parsimony methods are based on the whole genomic data and attempt to reconstruct the tree that minimises the number of substitutions on branches to produce the data (Fitch 1971). Parsimony methods are not currently

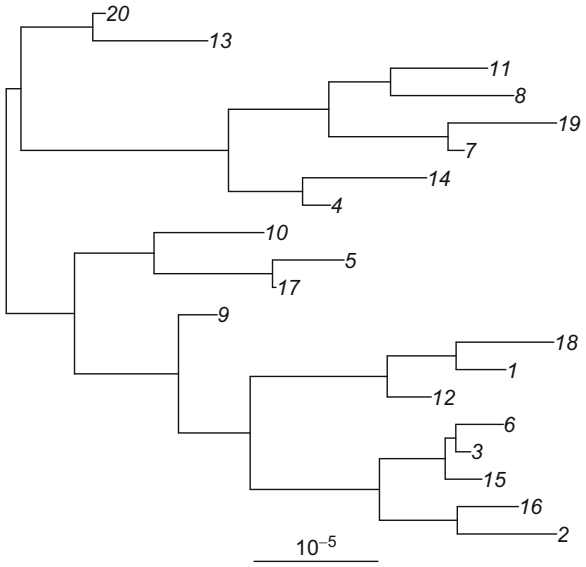

Fig. 1 Example of a phylogenetic tree. The x-axis represents evolutionary distance whereas the y-axis is arbitrary. It is important to 'read' the tree along the correct axis, for example genomes 4 and 10 appear next to other but are not especially closely related

frequently used to analyse microbial genomic data. Maximum likelihood techniques are based on an explicit probabilistic model of how substitutions accumulate on a tree, such that it is possible to define the likelihood, that is the probability of the genetic data given a tree. An efficient algorithm for computing the likelihood is the so-called pruning algorithm (Felsenstein 1981), which leaves the problem of exploring the space of all possible trees to find the one that maximises the likelihood. Powerful algorithms to do so have been developed that are implemented for example in the popular software phyml (Guindon et al. 2010), RAxML (Stamatakis 2006) and FastTree (Price et al. 2009, 2010). For any dataset with more than a handful of genomes, the number of possible trees is too large to allow a complete exploration of all trees, so that the analysis relies on heuristics which are not guaranteed to always return the best tree, but should still return one of the most likely trees.

Bayesian phylogenetic methods are based on an explicit evolutionary model like maximum likelihood but with two important differences. Firstly, the suitability of a tree is not measured in terms of the likelihood but of the posterior probability, which is the product of the likelihood and a prior probability. This term represents how appropriate a tree is deemed to be, only based on a tree model without reference to the genomic data. A prior tree model needs therefore to be specified, for example using the coalescent model (Kingman 1982), and this prior model can include parameters and be used to explore various evolutionary scenarios. Secondly, instead of finding a single maximising tree, the Bayesian approach returns a sample of trees that may have generated the data, also known as a posterior sample of trees. Comparisons between these trees can be performed to assess the statistical confidence in the phylogenetic reconstruction. In non-Bayesian phylogenetic techniques uncertainty measurement is usually achieved approximately and expensively using bootstrapping (Felsenstein 1985), but Bayesian phylogenetics provides a more natural way to do this. Popular software packages to perform Bayesian phylogenetics include MrBayes (Ronquist et al. 2012), RevBayes (Höhna et al. 2016), BEAST (Drummond and Rambaut 2007) and BEAST2 (Bouckaert et al. 2014).

3.2 Phylogenetics Accounting for Recombination

The phylogenetic techniques described in the previous section all assume that no recombination affected the data, so that a single tree applies for all sites. However, many microbes experience significant rates of recombination as they evolve. When this is the case, applying a method that assumes no recombination can lead to incorrect phylogenetic reconstructions (Schierup and Hein 2000; Hedge and Wilson 2014). A first sign of the effect of recombination can be obtained by estimating separate trees for different parts of the genome (for example, a tree for each gene). If recombination had not occurred, we would expect all such trees to look very similar, up to the randomness of the mutation process affecting each gene.

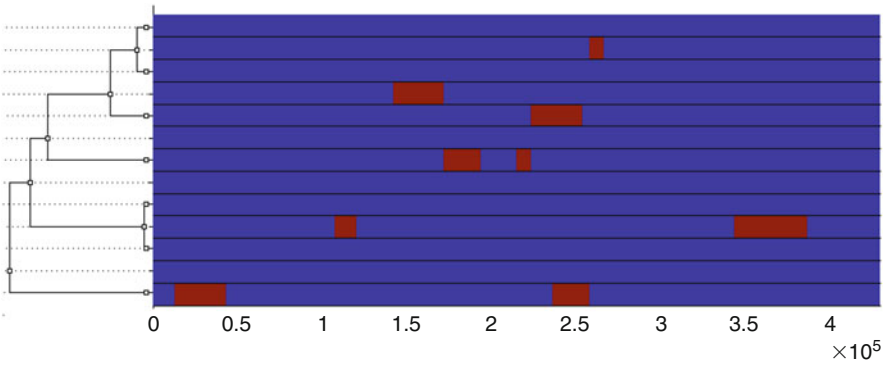

Fig. 2 Example of a phylogenetic tree with recombination events shown as a matrix on the right. To each terminal and internal branch of the tree corresponds a row of the matrix, with positions along the genome alignment shown on the x-axis of the matrix. For any given branch, unrecombined regions are shown in blue and recombined regions are shown in red

In some microbes, recombination happens exclusively as a gene conversion process, that is with a strong asymmetry between the two parents involved in recombination: the recipient cell contributes the vast majority of the resulting genome whereas the donor cell only contributes a short fragment. This is true of all bacterial species for example, irrespective of whether recombination was caused by conjugation, transduction or transformation (Didelot et al. 2010). In this case, recombination can be integrated into the phylogenetic tree reconstruction process by identifying the recombined fragments that happened on every branches of the phylogeny called clonal genealogy (Fig. 2). This clonal genealogy represents the ancestry process obtained by following the line of descent of the recipient at each recombination event, that is the line followed by the majority of the genetic material. A first software following this principle was ClonalFrame (Didelot and Falush 2007), which was originally designed for multilocus sequence typing data (Maiden et al. 1998) but can also work with limited (up to ~100) number of whole genomes, as was demonstrated for example by applications to *Escherichia coli* (Didelot et al. 2012b) and *Chlamydia trachomatis* (Joseph et al. 2011, 2012). For larger whole genome datasets, a newer version has been released which uses maximum likelihood optimisation techniques and is called ClonalFrameML (Didelot and Wilson 2015). ClonalFrameML has been applied for example to large genomic datasets of *Neisseria gonorrhoeae* (De Silva et al. 2016) and *Escherichia coli* (Ingle et al. 2016). A similar tool is Gubbins (Croucher et al. 2015) which operates through an iterative process of building a phylogenetic tree using standard recombination-unaware techniques, finding recombinant regions that do not fit the tree and repeating. Examples of application of Gubbins have been published on *Streptococcus pneumoniae* (Croucher et al. 2011) and *Chlamydia trachomatis* (Harris et al. 2012).

Instead or in addition to this gene conversion process, some microbes undergo recombination akin to crossing-over in higher organisms, that is where both parents

contribute large amounts of DNA. In this case, it is not possible to identify a recipient and donor for recombination events, and therefore there is not a defined clonal genealogy as above that can be targeted for phylogenetic reconstruction. This situation arises for many viruses, for example HIV. A solution is then to try and identify the breakpoints along the alignment where significant recombination events have occurred, and to reconstruct a separate phylogeny for each genomic region between two consecutive breakpoints. Computational software exploiting this idea include TOPALi (Milne et al. 2004, 2009), stepBrothers (Bloomquist et al. 2009), GARD (Pond et al. 2006) and RDP4 (Martin et al. 2015). A special recombination scenario occurs in the evolution of the influenza virus. The genome is made of eight segments, and recombination proceeds by replacement of whole segments, also known as reassortment. Techniques have therefore been developed to exploit this specific process, for example the GiRaF software (Nagarajan and Kingsford 2011) which reconstructs trees for each segment separately and considers the reassortment events that would be needed to reconcile them.

3.3 Non-phylogenetic Ancestry

A phylogeny is not always the best way to represent the ancestry of a sample of individuals. This is especially true for microbes that recombine a lot, as for example *Helicobacter pylori* in which 40% of genes can be affected by recombination within 3 years of within-host evolution (Kennemann et al. 2011). An alternative is to consider that there is a number (K) of underlying populations, with each individual either belonging to a population, or being a genetic mixture of the different populations (Fig. 3). One of the first algorithms to be based on this principle was STRUCTURE (Pritchard et al. 2000) and the linkage option (Falush et al. 2003) within it (as opposed to the non-admixture and admixture options) is especially useful to analyse sequence data since it models the correlation in the ancestry of sites near each other along the genome. For example, two sites next to each other have a high probability of having the same ancestry, since otherwise the boundary of a recombination event would have had to occur exactly between these two sites. The computational cost of running STRUCTURE does not scale well with the length and number of sequence being analysed though, and it is challenging to determine the number (K) of ancestral populations that should be considered in the model. Consequently, its current use in microbiology is limited to very specific situations, for example to quantify the admixture between the two bacterial species *Campylobacter jejuni* and *E. coli* (Sheppard et al. 2013a). Other softwares based on a similar population admixture principle include ADMIXTURE (Alexander et al. 2009) and BAPS (Tang et al. 2009) which is popular to determine population clusters amongst bacterial genomes, for example *Streptococcus pneumoniae* (Chewapreecha et al. 2014). Another non-phylogenetic approach is BratNextGen (Marttinen et al. 2012) which does not cluster individuals into populations as the previously mentioned software, but instead identifies the genomic fragments that

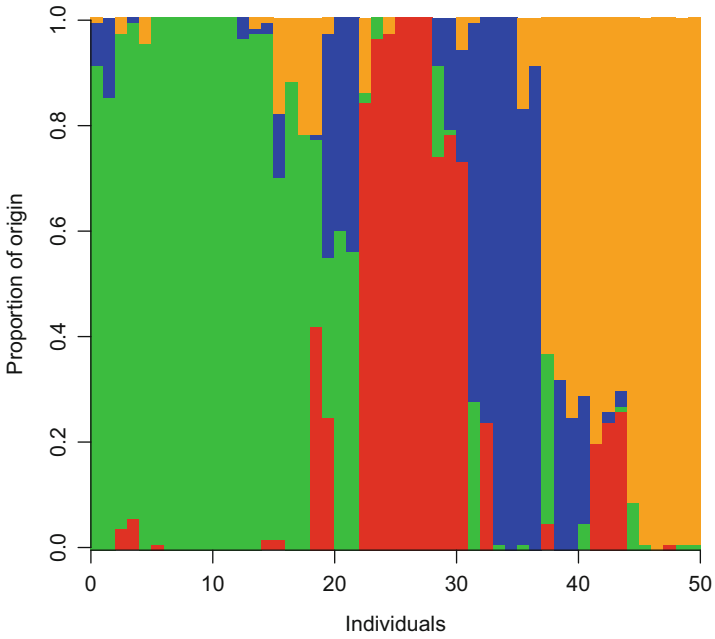

Fig. 3 Example of a barplot representation of population structure. The analysis includes 50 individuals shown on the x-axis and four populations have been detected, each of which corresponds to a colour (red, blue, green, orange). For each individual, the proportion of genomic material originating from each of the four populations is illustrated on the y-axis. The ordering of the individuals on the x-axis is arbitrary and often chosen to group together the individuals with similar profiles

are likely to have come from sources external to the population under consideration. BratNextGen is therefore usually applied to genomes from a single bacterial lineage to identify recombination events coming from other lineages, for example *Streptococcus pneumoniae* PMEN1 (Marttinen et al. 2012) or *Staphylococcus aureus* ST239 (Castillo-Ramírez et al. 2012).

FineStructure (Lawson et al. 2012) is another non-phylogenetic method to reconstruct the population structure. The algorithm proceeds in two steps. First each genome is considered in turn and reconstructed as a mosaic of all other genomes using a copying model (Li and Stephens 2003): each site is copied from one of the genome and copying occurs in blocks so that two neighbouring sites are likely to come from the same genome. The number of blocks copied by each genome from each other genome is then counted and summarised in a so-called co-ancestry matrix. A clustering method is then used to group together the individuals with similar co-ancestry rows into populations. Thus FineStructure reveals both the population of origin of each individual, and the fragments that have been imported from elsewhere, making it comparable to the previously mentioned linkage model of STRUCTURE (Falush et al. 2003). The computational cost of

running FineStructure is however much lower than that of running STRUCTURE, so that very large datasets can be analysed in a manner of hours. The problem of estimating the number of ancestral populations (K) is also resolved by the two-step approach. FineStructure was originally designed for human genetics, but has also proven useful in bacterial genomics, having been applied for example to *Helicobacter pylori* (Yahara et al. 2013), *Vibrio parahaemolyticus* (Cui et al. 2015) and *Myxococcus xanthus* (Wielgoss et al. 2016). An extension called orderedPainting has been developed specifically for detecting recombination hotspots in bacterial genomes (Yahara et al. 2014).

4 Integrating Temporal Data

4.1 Temporal Data in Microbial Genomics

The dates on which the microbes have been isolated are usually known, and it can often be interesting to integrate this information into the microbial genomic analysis. A first approach for doing so, which can be used in both phylogenetic and non-phylogenetic frameworks, is to simply annotate the reconstructed population ancestry with the dates, to see if some lineages or populations seem to have emerged more recently than others (for example, see Haase et al. 2014, Fig. 2b, d). In a phylogenetic framework, however, there is a more powerful approach available which is to try and reconstruct a timed tree (Fig. 4). In a timed tree, branch lengths are measured in a time unit (for example, days or years) rather than a genetic unit (for example, number of substitutions per site). Each tip represents a microbial genome and is aligned with its known date of isolate. Each internal node represents the most

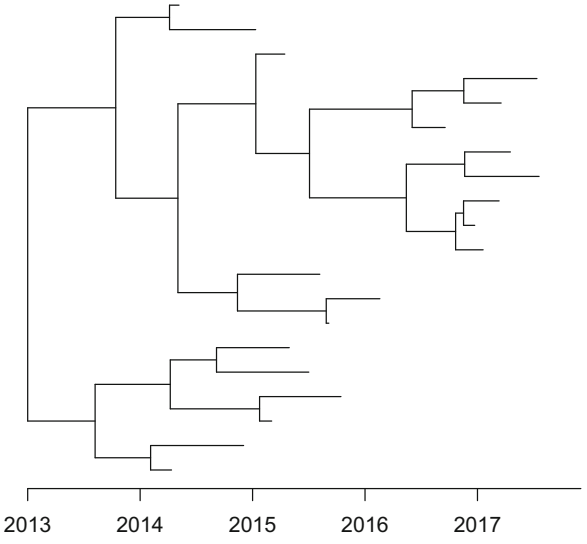

Fig. 4 Example of a timed tree. The interpretation is the same as for a standard phylogenetic tree, except that the time scale (x-axis) is measured in years rather than genetic distance. Each genome is aligned on the x-axis with its known date of isolation. Each internal node of the tree is aligned on the x-axis with the inferred date of existence of the last common ancestor of the genomes underneath

recent common ancestor between the set of genomes descended from the node, and is aligned with the date when it occurred, which is unknown but estimated by the phylogenetic procedure. In particular, the root of the tree represents the most recent common ancestor (MRCA) of the whole set of microbial genomes and it is aligned with the time to the most recent common ancestor (TMRCA) of the whole set. A timed tree therefore allows more natural interpretations to be drawn, especially when the research questions of interest are of an epidemiological or ecological nature, since the dating of all branches is included in the tree. Correctly reconstructing such a timed tree from a set of microbial genomes and associated isolation dates is therefore an important methodological concern.

4.2 Molecular Clock and Building a Timed Tree

Building a timed tree requires an estimate of the molecular clock rate, that is the rate at which substitutions are accumulated over time on genomes and measured for example in units of substitutions per year per site. Let us assume that there is such a rate and that it is relatively constant over the evolutionary history considered. This assumption is called the strict molecular clock assumption. Sometimes this rate has been estimated by previous studies and can be used directly to build the timed tree. For example, in a study of *Clostridium difficile*, genomes sampled longitudinally from the same hosts were compared to estimate the evolutionary clock rate, which was then used to produce timed trees (Didelot et al. 2012a). Otherwise, when the clock rate is unknown, it needs to be estimated from the data at hand. A simple approach for doing so is called root-to-tip method, where a non-timed phylogenetic tree is estimated, and a scatter plot is formed with a dot for each genome, the x-axis corresponding to the known isolation dates and the y-axis to the length of the path from root to the genome in the phylogeny (Fig. 5). If the strict clock assumption holds approximately, and that the range of sample dates is large enough relative to the age of the root, then a linear correlation should be found in this scatter plot. The slope of this linear regression is an estimate of the molecular clock rate, while the value on the x-axis at which the linear regression crosses the x-axis is an estimate of the age of the root of the phylogeny. This method was for example used in *Streptococcus pneumoniae* and showed much better results when based on a phylogeny that had been corrected for recombination compared to one that had not (Croucher et al. 2011). This root-to-tip method is useful to establish whether the temporal signal in the data is strong enough to consider applying the methods described below for reconstructing a timed tree. An implementation of the root-to-tip technique is provided by the software TempEst (Rambaut et al. 2016).

The most popular method to reconstruct a timed tree is that implemented in the softwares BEAST (Drummond et al. 2012) and BEAST2 (Bouckaert et al. 2014), relying on Bayesian statistics to jointly estimate the molecular clock, the timed tree and uncertainties around them. Reconstructing timed trees using BEAST has been especially popular for analysing viral genetic data, for example in influenza (Smith et al. 2009), HIV (Worobey et al. 2008) and Ebola (Gire et al. 2014), but more

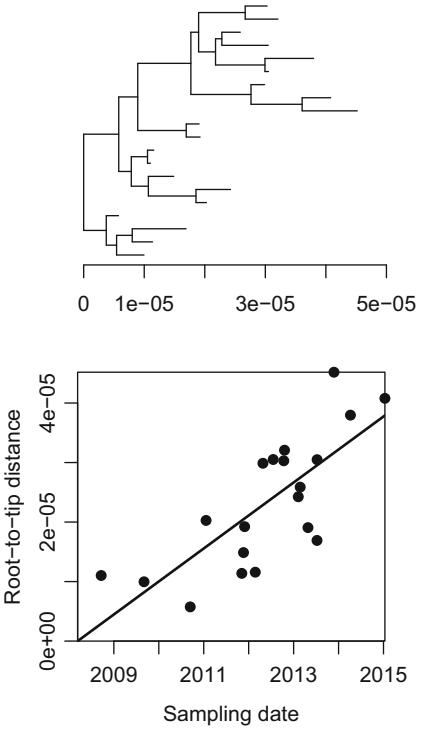

Fig. 5 Example of application of the root-to-tip method. The top panel shows the phylogenetic tree reconstructed for the genomes of interest. On the bottom panel, there is a dot for each of these genomes, with the x-axis representing the known date of isolation of the genome and the y-axis representing the length of the path from root to tip in the phylogenetic tree. A linear regression can then be attempted on the scatter plot, which if statistically well supported can be used to estimate both the molecular clock rate (slope of the regression) and the time of the most recent common ancestor for the whole set of genomes (intersect of the linear regression with the x-axis, here 2008)

recently has also gained in popularity for bacterial genomics (Biek et al. 2015), for example in the study of *Yersinia pestis* (Cui et al. 2013), *Shigella sonnei* (Holt et al. 2012) and *Escherichia coli* (Stoesser et al. 2016). BEAST also implements options to use instead of the strict molecular clock described so far, a relaxed molecular clock where the rate of evolution is allowed to vary to some extent between the different branches of the tree (Drummond et al. 2006; Drummond and Suchard 2010). An alternative to BEAST is LSD (To et al. 2016) which is faster and able to deal with larger datasets as was demonstrated for example recently in an analysis of thousands of simulated HIV genomes (Ratmann et al. 2017).

4.3 Phylodynamics

Past changes in population size affect what a timed genealogy is likely to look like (Griffiths and Tavare 1994). For example, if the population size has been increasing significantly, it will result in longer terminal branches and shorter internal branches compared to a tree under a constant or declining population size. It is also possible to turn this stochastic relationship around, meaning that a reconstructed timed phylogeny is informative about past population size dynamics. Phylodynamics is

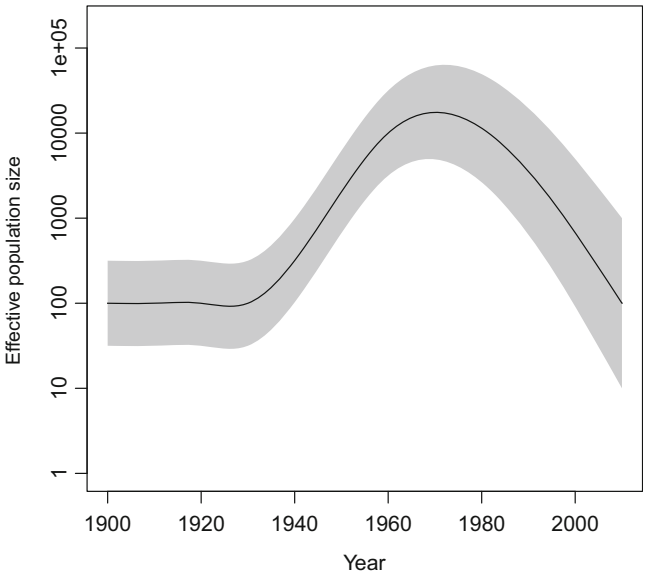

Fig. 6 Example of a skyline plot. The black line indicates the estimated population size over time, with the grey shading representing the 95% credibility interval. Here we see that the population size was stable from 1900 until 1940, increased significantly up until 1965 after which it started to decline back to its original level

the branch of phylogenetics that exploits this property. Following their implementation into the BEAST framework, starting with the Bayesian skyline plot (Drummond et al. 2005), these techniques have become increasingly popular to analyse microbial genomic data. The typical result is a plot with time on the x-axis and the effective population size on the y-axis (often measured on a log scale), with a line indicating the mean estimated population size variations and shading representing the 95% credibility interval over time (Fig. 6). Phylodynamics is very popular to investigate viral population size dynamics, for example in an analysis of rabies in North American raccoons where the skyline plot is in good agreement with epidemiological information about the spread of the disease (Biek et al. 2007). It is also sometimes used in bacterial genomics, for example in a study of the emergence of *Staphylococcus aureus* ST225 in Germany and the Czech Republic (Nübel et al. 2010).

5 Integrating Spatial Data

5.1 Using a Descriptive Approach

When the spatial origin of the genomes is known and varied, using this information in the context of a microbial population genomic analysis can help to reveal the

geographical structuring of the population and the potential occurrence of migrations between locations. The spatial data used for such a phylogeographic analysis can occur at any scale, including between patches of land separated by only a few centimetres (Wielgoss et al. 2016), between different body parts within a single host (Didelot et al. 2016), between different regions of a single country or between different countries throughout the world (Croucher and Didelot 2015).

The simplest approach to investigate the geographical pattern of the origins of the microbial genomes is to plot the geographical data side-by-side with the results of the analysis of population ancestry. If a non-phylogenetic, clustering method was used for the analysis of population ancestry, then the distributions of geographical origins can be compared between inferred clusters. If a phylogenetic method was used, the leaves of the tree can be annotated according to spatial origin, for example by using a different colour for each location. For example, these two types of annotations (non-phylogenetic and phylogenetic) were both used in a genomic analysis of *Streptococcus suis* (Weinert et al. 2015) in Figs. 1c and 5, respectively. This purely descriptive approach can already reveal interesting features and, in the phylogenetic context, the extent to which genomes from each location form clusters in the tree is noteworthy. Such clustering is indicative of the strength of the geographical structure and exceptions where a genome falls into the "wrong" cluster can represent recent migrations, as was shown for example in a global genomic analysis in *Staphylococcus aureus* ST239 (Harris et al. 2010). Likewise, when the aim is to investigate the source of an isolate, simply looking at the origins of its nearest relatives can be highly suggestive, as was used for example to uncover the South-East Asian origin of the 2010 Haiti cholera outbreak (Chin et al. 2011). The Microreact web interface (Argimón et al. 2016) provides a user-friendly way of studying side-by-side the origin of isolates on a map and their genomic relationships, including the ability to interactively explore subsets of isolates defined by geographical or genomic criteria.

5.2 Using an Inferential Approach

A natural next step beyond annotating the leaves of a tree with spatial sources is to try to annotate the internal nodes or branches (Fig. 7). However, doing this requires an algorithm to infer the ancestral locations since this is only known about the leaves. The most widespread approach for doing so is to consider the location as a discrete trait that evolves along the branches of the tree, with mutations of the discrete trait corresponding to migrations from one location to another. Migrations occur according to an unknown matrix of rates from any location to any other, which may be constrained to reduce the number of parameters to estimate, for example by considering that migration from location A to location B happen at the same rate as from location B to location A, so that the migration rate matrix becomes symmetric. Joint inference of the migration matrix and ancestral locations can be performed under such a model using ancestral state reconstruction

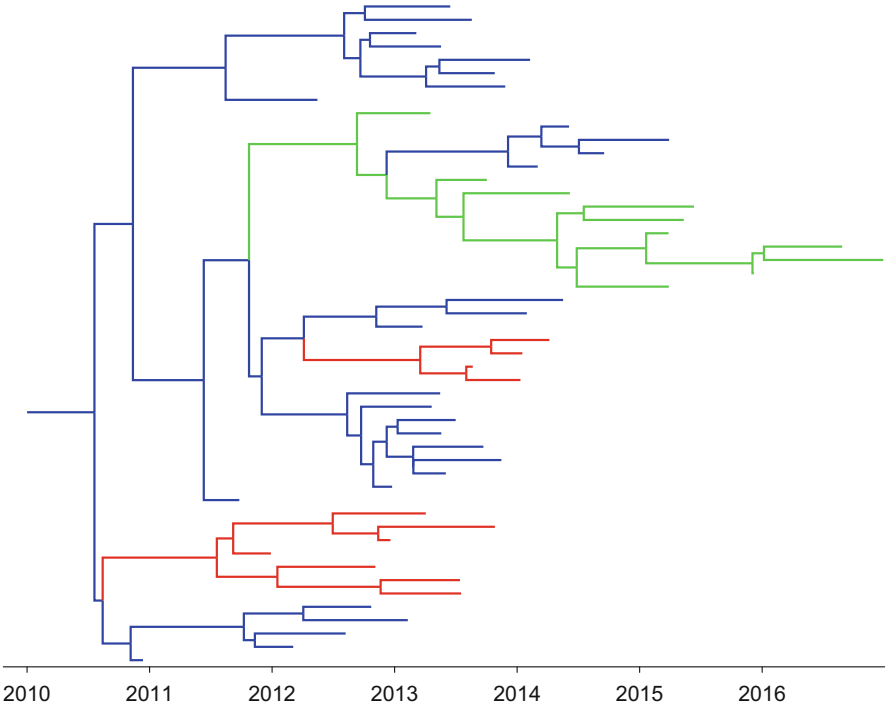

Fig. 7 Example of tree coloured by geographical location. A colour is assigned to each of the locations (here for example three countries are shown in red, green and blue). The location of origin of each genome is known and shown by colouring the corresponding terminal branch with the appropriate colour. The location of ancestors is not known but can be inferred using algorithms as described in the main text and this can then be shown by colouring internal branches of the tree accordingly

techniques (Joy et al. 2016). For example, the ancestral location of *Shigella sonnei* lineages was reconstructed (Holt et al. 2012) by maximum likelihood estimation using the ace command from the R package ape (Paradis et al. 2004). Once the ancestral locations have been reconstructed, the full history of past migrations is revealed since changes in location along a branch or from one branch to its descendent branch can be interpreted as a migration from one location to another. When combined with temporal information (see previous section), this approach can reveal the spatio-temporal spread of a microbe, for example the global spread of the current pandemic of cholera in three waves that all originated from South-East Asia (Mutreja et al. 2011; Didelot et al. 2015).

Phylogeographic analysis can also be performed within the BEAST and BEAST2 frameworks (Drummond et al. 2012; Bouckaert et al. 2014), either using the discrete trait modelling approach described above (Lemey et al. 2009) or a continuous space version (Lemey et al. 2010). The latter has the advantage to analyse the ancestral locations at the same time as the phylogenetic space is being

explored, so that phylogenetic uncertainty is accounted for in the phylogeographic analysis. This technique was originally applied to Avian Influenza A H5N1 (Lemey et al. 2009) and rabies (Lemey et al. 2010) and has since become very popular mostly for viral phylogeography studies (Bloomquist et al. 2010) but also to investigate bacterial phylogeography such as *Mycobacterium tuberculosis* (Comas et al. 2013) and *Clostridium difficile* (He et al. 2013). Powerful interactive visualisation techniques have also been developed to explore the ancestral reconstructions output by these analytical methods (Bielejec et al. 2011, 2016). Within BEAST2 (Bouckaert et al. 2014) a separate algorithm called BASTA has recently been developed which is based on an approximation of the structured coalescent and can lead to more accurate ancestral reconstructions, especially when sampling is highly biased between locations (De Maio et al. 2015).

6 Integrating Other Types of Data

6.1 Application of Ancestral State Reconstruction

Non-genomic metadata can be integrated with phylogeny to provide insight into the evolutionary history of populations. When performing a microbial population genomics study, there are often additional non-genomic metadata that it can be interesting to integrate into the analysis to investigate their relationship with the evolutionary history of the population. The last two sections described specifically the case of temporal and spatial data, and this section discusses the use of other types of data. Depending on the system under study, this metadata may include virulence measurements, antimicrobial resistance profiles, host species of origin, tissue of origin, conditions of isolation, results of in vitro experiments, etc.

Many of the methods described in the previous section for the analysis of spatial data can be applied to other types of metadata, because they are based on models of discrete or continuous trait evolution that are not specific to phylogeographic analysis. The evolutionary history of traits of interest can thus be revealed in the form of changes in the metadata value along branches of the tree (when working in a phylogenetic framework), or significant differences between populations (when working in a clustering framework). For example, maximum likelihood estimation of ancestral state, as implemented for instance in the ace command of the R package ape (Paradis et al. 2004), was used to reconstruct the evolutionary history of pathogenicity in *Clostridium difficile* (Dingle et al. 2014). Likewise, even though the discrete trait analysis methodology implemented in BEAST (Lemey et al. 2009) was originally developed with phylogeography in mind, it has since been applied to other non-spatial traits. In bacterial population genomics, examples include studies of host species in *Campylobacter jejuni* (Dearlove et al. 2015) and host sexual orientation in *Neisseria gonorrhoeae* (Grad et al. 2014). In viral population genomics, examples include studies of host species in rabies (Faria et al. 2013) and antigenic diversity in influenza (Zinder et al. 2013).