Multilocus Sequence Typing of Pathogens: Methods, Analyses, and Applications

16

M. Pérez-Losada [1,2,3], *M. Arenas* [4,5], *E. Castro-Nallar* [6]
[1]George Washington University, Ashburn, VA, United States; [2]Universidade do Porto, Vairão, Portugal; [3]Children's National Medical Center, Washington, DC, United States; [4]University of Porto, Porto, Portugal; [5]Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal; [6]Universidad Andrés Bello, Santiago, Chile

1. Introduction

The ability to accurately distinguish between strains of infectious pathogens is crucial for efficient epidemiological and surveillance analysis, studying microbial population structure and dynamics and, ultimately, developing improved public health control strategies.[1] To further such general goals, several molecular typing methods have been proposed that can identify isolates worldwide (global epidemiology) and/or in localized disease outbreaks (local epidemiology); see Foley for a review.[2] Nonetheless, since 1998, the established standard for molecular typing is multilocus sequence typing[3] (MLST). MLST was built on the well-established population genetic concepts and methods of the multilocus enzyme electrophoresis (MLEE) technique, but provides significant advantages over this and other typing approaches (see Section 4 for advantages and caveats). MLST examines nucleotide variation in sequences of internal fragments of usually seven housekeeping genes: that is, those encoding fundamental metabolic functions (see Section 2 for molecular design and development of MLST). For each gene, the different sequences present within a species are assigned as distinct alleles and, for each isolate, the alleles at each of the seven loci define the allelic profile or sequence type (ST). Each isolate is therefore unambiguously characterized by a series of seven integers, which correspond to the alleles at the seven housekeeping loci. Most bacterial species have sufficient variation within housekeeping genes to provide many alleles per locus, allowing billions of distinct allelic profiles to be distinguished using just seven loci. Alternatively, isolate identification and tracking can be performed using the nucleotide data directly, although this approach is more frequently used for population studies (see Section 5 for methods of analyses).

MLST is widely used for molecular typing.[4–7] Numerous examples exist of their use for describing the population structure of pathogens, vaccine studies, tracking transmission of epidemic strains, and identifying species and virulent strains associated with disease (see Section 6 for applications). This was made possible by three

Genetics and Evolution of Infectious Diseases. http://dx.doi.org/10.1016/B978-0-12-799942-5.00016-0

improvements in molecular microbiology[4] involving: (1) bacterial evolution and population biology knowledge (discussed later); (2) high-throughput nucleotide sequencing (see Section 2 for molecular basis); and (3) internet databases (see Section 3). The bacterial population studies undertaken from the 1980s onward were central to the development of MLST. Those studies showed that genetic exchange among bacteria was more common than previously thought, leading to a reassessment of the role of sexual processes in the structuring of bacterial populations. Using sequence data, it has been shown that recombination (mosaic genes) was frequent not only in genes under diversifying selection (e.g., antigen-encoding and antibiotic-resistant determinant genes), but also in genes under purifying selection (housekeeping genes). This suggested that the clonal model (variation can only arise by mutation) was not universal and led to the proposal of new nonclonal or panmictic (variation is mainly generated by recombination) and partially clonal models of bacterial population structure. Consequently, typing methods needed to accommodate a broader spectrum of population structures and be able to distinguish among them, hence providing not only discriminatory power but also information about the clonal structure of the organism under study. Therefore, only molecular techniques that can contrast results across independent markers (such as MLST) would be adequate for bacterial typing and population genetic inferences.

In the following sections, we describe in more detail all the epigraphs mentioned in this introduction. We also refer the reader to other reviews on MLST for complementary information.[1,4–6,8–13]

2. Molecular Design and Development of Multilocus Sequence Typing

The principal element in the design of an MLST scheme is the choice of genetic loci. The selection and number of loci is based on principle, precedent, and practice. Since MLST was developed as an updated version of MLEE, which indexes variation of multiple core metabolic or housekeeping genes at the protein level, the selected loci typically correspond to housekeeping genes encoding proteins for core metabolic functions. Furthermore, housekeeping genes are expected to be somewhat conserved and vertically transmitted and thus should reveal genetic relationships among strains without concern for the influence of host or environmental factors. For instance, such influences might occur when genes encoding hypervariable surface proteins are subject to immune-driven diversifying selection or genes under antibiotic selection. The genes should be physically spaced around the genome in order to minimize genetic linkage of loci.

As a matter of principle and practicality, multiple loci of sufficient length need to be surveyed in order to provide a high level of discrimination. The first MLST scheme was designed by Maiden and colleagues[3] and included six, later expanded to seven, loci. Most investigators have followed this precedent and developed schemes of seven loci. The length of nucleotide sequence amplified for each locus is generally in the

range of 400−600 bp and is determined largely by the parameters of automated sequencing instruments available at the time the first MLST scheme was developed in 1998. Most MLST nucleotide sequence data are generated by the Sanger sequencing method, however, high-throughput technologies such as pyrosequencing,[10,14] sequencing-by-synthesis, and single-molecule sequencing[5,15] will likely be the methods of choice in the future for both targeted-amplicon and whole-genome sequencing. Those technologies are capable of generating accurate read lengths of ~500 bp to 10 kb (PacBio RS II and Sequel systems) and up to 25−50 million paired-end reads (Illumina MiniSeq/MiSeq platforms) per run. Moreover, the design of barcoded primers allows simultaneous and efficient sequencing of homologous products from hundreds of samples in the same run[16]; see also www.pacb.com and Chen et al.[15]

The development of a new MLST scheme from scratch involves four initial steps (Table 16.1): (1) identification of loci, (2) PCR primer design, (3) survey of a small number of representative strains, and (4) analysis of nucleotide sequence data to

Table 16.1 Stages in the Design of an MLST Scheme

Actions	Criteria
Analyze reference genome to identify 12−18 candidate loci	• Single-copy gene • Putative core housekeeping gene Genes evenly spaced in genome
Design nested PCRs using primer select software	• Outer PCR product about 1000−1500 bp • Inner PCR product about 400−600 bp
Select 20−25 representative strains	• Isolated in different years and different geographic sites • No known epidemiological linkage by transmission or shared phenotypic characteristics
Perform nested PCRs of the 20−25 strains and redesign primers as needed	
Analyze nucleotide sequence data	• Rank loci by level of nucleotide polymorphisms and select 7−9 loci with high level of polymorphism
Select 75−100 strains using the previously mentioned criteria, type using the 7−9 loci, and perform analysis	• Confirm loci are under purifying selection • Assign each unique sequence an allele number • Assign each isolate an ST • The greater the number of STs, the greater the discriminatory power of the MLST

establish neutral evolution of loci and level of strain discrimination. For many bacterial species, the selection of loci is greatly aided by the availability of annotated whole genomes, which allows ready identification of housekeeping genes and their physical location in the genome. An absolute requirement for loci included in an MLST scheme is that there is only a single copy of the gene in the genome. It is advisable to choose more than seven loci because not all loci will pass subsequent tests of utility, and typically 12−18 loci are selected for subsequent tests. As much as possible, the loci should be evenly spaced across the genome and certainly separated by several tens of thousands of base pairs, although no rules allow a precise estimate of the maximum size of bacterial genomic fragments that can undergo recombination. The physical location of loci within genomes may differ among strains, so use of a single reference strain, which is often all that is available, is at best an approximation. The design of primers is greatly assisted by the availability of open access and commercial software for primer design but ultimately depends on trial and error.[17,18] Most MLST schemes use a nested PCR design both to increase sensitivity for samples with a low bacterial DNA copy number and, more importantly, to provide a high-quantity and high-quality PCR product for sequencing. The initial evaluation of candidate loci is most easily accomplished with a small number of strains (20−25), and the strains should not be epidemiologically linked or share defining characteristics, such as antibiotic resistance, that might lead to over-sampling of a clonal population. Temporally and geographically separated strains provide one likely basis for accomplishing this goal. The data from this small set of strains should allow the stratification of loci on the basis of efficiency of detection by nested PCR and level of genetic variation. They also provide the opportunity to optimize primer design.

At least 7−9 loci that could be amplified from all test strains and showed a reasonably high level of genetic diversity[5,19] should then be evaluated with a larger data set of 70−100 strains to accomplish the initial data analysis for evolutionary neutrality and level of strain discrimination. The same rules for selection of strains apply here as mentioned earlier. A representative collection of strains should be used, but in practice it is only possible to avoid obvious pitfalls, such as selecting strains from a known outbreak. The purpose of the initial analysis of MLST data is to confirm that the chosen loci are under purifying selection, to assess the level of polymorphism at each locus, and to determine whether a sufficient level of discrimination is achieved for epidemiological studies. The number of unique nucleotide sequences among the 70−100 strains tested establishes the level of polymorphism, and alleles that are the most polymorphic will provide the greatest degree of discrimination among strains. While low levels of polymorphism are a reason to reject an allele for inclusion in an MLST scheme because they will provide little discriminatory power, the seven most polymorphic alleles are not necessarily the best choice. Ideally, all seven loci will contribute equally to the discriminatory power of the method and a very high level of variation may be indicative of diversifying selection pressure. On the other hand, evolutionary neutrality is a desirable, but not absolutely necessary, characteristic of loci used in an MLST typing scheme. In fact, most other methods for strain typing use highly polymorphic loci, which are often known to be subject to selection pressure. If one or more of the initially selected loci fail the test of neutrality, or no

combination of 6−7 loci provides sufficient strain discrimination, other loci surveyed in the test set can be evaluated with the larger data set, and a new 6−7 loci MLST scheme can be designed. Finding the right balance in terms of efficiency of PCR amplification, locus neutrality, strain discrimination, and comparability of polymorphisms across loci is ultimately a matter of judgment rather than the application of precise rules.

Once candidate loci have been chosen and the MLST scheme defined, application of the method in the context of epidemiological studies will establish its reliability in typing large numbers of diverse strains and its ability to provide sufficient strain discrimination to address epidemiological questions of interest. For strains that cannot be typed using the initial PCR primers, it is generally easy to design new primers. Although the choice of loci used in the MLST scheme could be modified as more strains are typed (e.g., to increase discrimination), one of the strengths of MLST as a typing method would be sacrificed; namely, the comparability of data generated over time and by multiple investigators. Because the sequence type or ST is defined by the set of distinctly numbered alleles at the seven loci, changing loci would result in new STs that could not be directly compared to STs defined using the previous MLST scheme. In that regard, using in silico MLST approaches based on whole-genome data allows us to compare different typing schemes for the same group or even integrate genomic inferences with information-rich MLST databases.[20−23]

If an epidemiological study requires discrimination of closely related strains, as may be necessary to examine short-term transmission of antibiotic-resistant isolates, rather than add to or change the loci in an MLST scheme, a better strategy is to supplement MLST with additional highly polymorphic markers, such as genes encoding antigens, cell surface proteins, ribosomal genes, or tandem repeats.[11,19,24−26]

Over the last few years, other typing approaches have been developed based on similar principles as MLST. Multilocus Variable number of tandem repeats Analysis (MLVA) uses polymorphic repeated sequences (VNTR) instead of housekeeping genes. Comparative studies between MLVA and MLST have yielded similar results, for example, van Cuyck et al.,[27] and in recently originated species, the MLVA approach may have higher discriminatory power.[28] Similarly, the Ribosomal Multilocus Sequence Typing method (rMLST) has also been proposed to index the molecular variation of 53 genes encoding bacterial ribosome protein subunits.[11] This method pursues the integration of a taxonomic and typing method in a similar curated MLST scheme. Although more expensive, the rMLST is likely to provide better resolution than previous methodologies. Likewise, core-genome (cg) MLST has been developed to overcome lack of resolution of MLST schemes of certain taxa. By collecting a sample of genome sequences representing extant diversity, the cgMLST scheme uses >1000 genes to create sequence types that provide increased resolution for clonal populations of bacteria.[29] Finally, in order to achieve even greater resolution, other approaches have been developed based on core/accessory genes or distributed genes among bacterial species that have the same MLST profile.[30,31] This new approach could skip the laborious and time-consuming steps needed to develop bacteria-specific MLST schemes.

3. Multilocus Sequence Typing Databases

One of the goals of the MLST approach was the development of online platforms containing MLST databases to which public health officials and researchers could both have access and contribute; and from which clinical, epidemiological and population studies could benefit.[3,4,8] The first MLST websites were based on single databases implemented in the MLSTdB software[32]; but as MLST schemes began to expand, several limitations became apparent: redundant information (each record contained the ST designation and the allelic profile), isolate bias (single databases were dominated by specific studies), and access (all databases were stored at a single location). To overcome these limitations, a new network-based database software, MLSTdBNet,[33] was developed and implemented on the PubMLST site (http://pubmlst.org/). This site is served by two databases: (1) a profiles database that contains the sequences of each MLST allele for each locus linked to an allele number, and (2) an allelic profiles database with their ST designations. The profile database can then serve other isolate databases. For each scheme on the PubMLST site there is a PubMLST isolate database that aims to include at least one isolate for each ST. MLST databases are hence different from other depository databases, such as GenBank, not only in organization but also in that they are actively curated for accuracy. It is important to highlight that MLST databases do not embody the global diversity of an organism but the extent of its diversity at the time they are accessed. Moreover, stored data is unstructured and does not necessarily represent natural populations either. As high-throughput sequencing becomes more affordable, PubMLST is increasingly including whole-genome sequences, for example, BIGSdb.[34]

Several other websites are accessible through the PubMLST site. The PubMed (NCBI) is linked to PubMLST databases, so original publications describing MLST schemes can be retrieved. The AgdbNet—antigen sequence database software for bacterial typing[35]—is also integrated into the system. Other websites are available for the storage and access of MLST data. At the time of writing, 93 MLST schemes (82 for bacteria, 9 for eukaryotes, and 1 each for plasmids and bacteriophage) could be accessed via the PubMLST site. The PubMLST primary site is also mirrored in four locations, three in UK and one in Pittsburg (USA). This provides access to MLST data globally and assures that databases are stored in multiple locations. A detailed description of the MLST databases, their structure, and most of the published MLST schemes can be found in Maiden.[4]

Other websites (www.spatialepidemiology.net/ and beta.mlst.net/Instructions/mlstmaps.html) have also been developed that incorporate geospatial information in bacterial epidemiological studies. Those websites provide precise locality data related to strain distribution and a map-based interface for displaying and analyzing epidemiological information. Moreover, the portal www.eMLSA.net enables species identification by means of a taxonomic platform. The integration of genomic and epidemiological data together with geographic information through MLST databases will greatly improve our ability to track and prevent infectious pathogens and associated diseases.

4. Advantages and Disadvantages of Multilocus Sequence Typing

As the number of schemes available has increased, MLST has become the most commonly used method of pathogen typing. In comparison to older methods (serotyping; MLEE), the use of genetic variation gives MLST the advantage of producing variable data (more resolution) that are universally comparable (within schemes), easily validated, and readily shared across laboratories. The use of generic sequencing technology makes MLST a broadly applicable methodology that can be fully automated and scalable from single isolates to thousands of samples. Because the materials needed for MLST analysis—DNA or dead cells—are easily transported among laboratories without the problems associated with infective materials, both the biological samples and the resulting data are highly portable. Furthermore, the use of online electronic databases (see Section 3) to store and curate MLST schemes makes them a globally accessible resource.

MLST targets variation at multiple housekeeping loci. The number of loci that need to be evaluated to confidently assign an ST has been minimized to reduce the expense and time required for characterization, with most studies using 6−10 loci. If performed manually, evaluating even this many loci can be time consuming. However, fully automated systems, for example, robotics[36] provide a high-throughput pipeline for data collection that can run large volumes of samples with increased reliability. Likewise, commercial solutions, such as Ion Torrent AmpliSeq panels targeting MLST schemes, can reduce costs down to cents per marker (www.ampliseq.com). As sequencing technology progresses, we expect the cost of automation to decrease, so data interpretation, rather than data generation, will be the likely limiting factor in our understanding of pathogen population dynamics.

By focusing on sequence variation, MLST provides a highly replicable and reproducible typing method. Additionally, the focus on housekeeping genes provides significant amounts of genetic data that can be used to calculate pathogen population genetic parameters (see Section 5) at both local and global scales. Those parameters can be then used to construct more sophisticated models of pathogen evolution and epidemiology that will improve our understanding of how to control the spread of disease. However, there is no single set of universal housekeeping genes that can be used for all pathogens as the recombination rates, substitution rates, and levels of selection vary across loci and species.[13] Therefore, a unique set of loci must be identified for each novel, untyped pathogen under study. The rapid increase of available microbial genomes will make data mining for housekeeping genes more feasible, reducing the time and cost required for constructing new MLST schemes.

Currently, the main drawback of the MLST method is that the selection of housekeeping loci requires reference genomes.[37] Moreover, not all pathogens are suitable for MLST methods. Some pathogens (e.g., *Mycobacterium tuberculosis* and *Yersinia pestis*) exhibit very little variation throughout their entire genome, most likely representing "evolutionarily young" pathogens that have not yet accumulated sufficient genetic variation to differentiate strains. For typing these pathogens, more rapidly

evolving loci (e.g., insertion sequences or antibiotic-resistance determinants) or more markers (genome-wide SNPs) are needed. Conversely, some bacterial genomes have accumulated so much variation that MLST housekeeping genes do not provide adequate information for typing. As we advance MLST schemes in the postgenomic era, we should be able to combine information-rich and widely adopted schemes with cost-effective whole-genome sequencing.

5. Analytical Approaches

There are two basic strategies to the analysis of MLST data (Fig. 16.1), one relies on allele and ST designations to estimate relatedness among isolates (*allele-based methods*), and so ignores the number of nucleotide differences between alleles; and the other relies on nucleotide sequences directly to estimate relatedness and population parameters (*nucleotide-based methods*). The allele-based approach has been adopted from the analysis of MLEE data and so methods based on this strategy were the first applied to the analysis of MLST data.[3,38] The allele-based approach is thought to work well in nonclonal organisms (e.g., *Helicobacter pylori*), while nucleotide-based

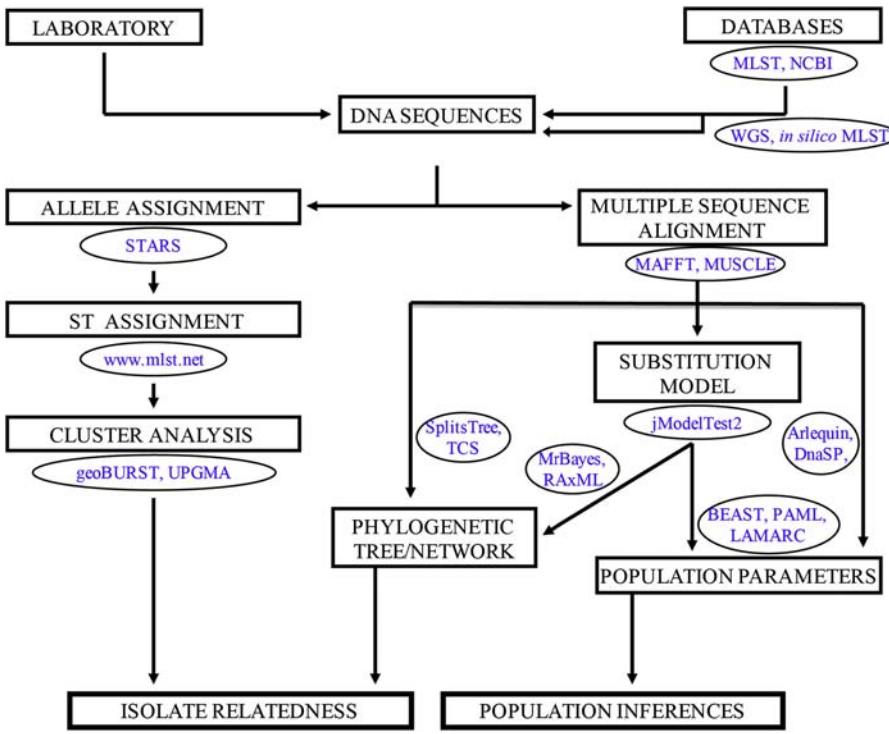

Figure 16.1 Pipeline showing data and tasks (*boxes*) and databases and computer programs (*circles*) commonly used in the analysis of MLST data.

approaches are preferable for clonal organisms (e.g., *Escherichia coli*) since the former are likely misleading.[4] But in practice, most microbes show some degree of clonality (clonal complex) in their populations; hence, in our opinion, both types of analyses should be conducted in population and epidemiological studies, for example, Loubna et al.[39] In this section, we present a brief description of some of the most commonly used approaches for analyzing MLST data. We refer the reader to previous reviews for a more detailed description.[9]

5.1 Allele-Based Methods

Since alleles are the unit of analysis, all these methods first require assigning an allele number to each DNA sequence from each locus. This is done by matching our sequences against those stored in public MLST databases (see Section 3). If no match is found, a new number is assigned in order of discovery. Several computational programs have been developed for this task, although Sequence Typing Analysis and Retrieval System (STARS) seems to be very functional and widely popular.[9] The STARS interface was specifically designed for typing and allows the assembly of large number of sequences at once.

Once alleles have been assigned, data are entered in the MLST websites to acquire an ST profile. At this point, exploratory analysis (e.g., allele and profile frequencies, polymorphism estimates, and codon usage) of the data can be performed. The software package Sequence Type Analysis and Recombinational Tests (START2) can perform all these tasks.[40] Relatedness among STs can be then displayed using methods of cluster reconstruction, such as the Based Upon Related Sequences Types (eBURST) approach and the simple Unweighted Pair Group Method with Arithmetic Mean (UPGMA). eBURST[41] is based on a simple model of clonal expansion and diversification. It first identifies mutually exclusive groups of related STs and attempts to identify the founding ST of each group. Bootstrap estimates are also calculated to assess confidence in the groupings. The algorithm then predicts the descent from the predicted founding ST to the other STs in the group, displaying the output as a radial diagram, centered on the predicted founding ST. A globally optimized version (goeBURST) is also available that identifies alternative patterns of descent using a graphic matroid approach.[42] In 2012, a new approach (PHYLOViZ) was released for microbial epidemiological and population analysis that allows for the integration of allelic profiles from MLST or MLVA methods (although Single Nucleotide Polymorphism data can also be included) and associated epidemiological data.[43] PHYLOViZ uses goeBURST for representing the possible evolutionary relationships between strains.

The traditional UPGMA method relies on a matrix of distances to estimate isolate relatedness. Distances are calculated for each pair of STs based on the number of allele differences, and groups are then sequentially clustered in order of similarity (i.e., allelic matches). Additional distance and parsimony methods have been proposed to estimate relatedness based on allele frequencies, but note that distance methods generally outperform parsimony methods.[44]

Allele-based methods have the advantage of simplicity and speed, which are crucial for efficient epidemiological surveillance and public health management, but disregard much of the evolutionary information contained at the nucleotide level. A larger and more sophisticated plethora of nucleotide-based methods exist to estimate isolate relationships and key population parameters.

5.2 Nucleotide-Based Methods

Any analysis of nucleotide data usually begins with a multiple sequence alignment (MSA) (i.e., estimation of the homologous nucleotide sites). Since the loci used for MLST usually evolve very slowly and code for proteins, this step becomes trivial, particularly at the amino acid level. If needed, several fast and accurate iterative aligning strategies are implemented in MAFFT[45] and MUSCLE.[46]

Once an alignment has been generated, we have to determine the model of evolution that fits the data the best. Model choice is a critical issue and the implemented model (or lack thereof) will affect all subsequent phylogenetic[47] and population analyses (following two sections). This issue is usually assessed within a phylogenetic framework, see Posada et al.[48] Since mid-1990s substitution models have increased in complexity, as parameters reflecting new information on nucleotide substitution processes are added to candidate models.[49] Furthermore, model selection can consider confidence sets of models (model averaging).[48] Several criteria have been proposed for choosing models, such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Decision Theory (DT), and Hierarchical Likelihood Ratio Test (hLRT).[50] Although AIC is the most broadly used method for evaluating model fit, BIC and DT should be preferred.[51] These strategies are implemented in the well-established program jModelTest2.[50]

5.2.1 Phylogenetic Relatedness

Phylogenetic reconstruction methods can be divided into two types, those that proceed algorithmically through distances, for example, UPGMA and neighbor joining (NJ) and those based on optimality criteria. Here, we focus on those that implement maximum likelihood and Bayesian optimality criteria and allow for the implementation of multiple data partitions each under its best-fit model. We find this feature particularly important for analyzing MLST data.

Maximum likelihood (ML) inference attempts to identify the topology that explains the evolution of a set of aligned sequences under a given substitution model of evolution with the greatest likelihood.[52] RAxML[53] implements the ML criterion efficiently and accurately and can handle large data sets of >1000 sequences with >20 kb.[54] Confidence in the estimated relationships (i.e., clade support) is usually assessed using a nonparametric bootstrap procedure,[55] which must be repeated >1000 times to achieve reasonable precision. RAxML can also rapidly estimate bootstrap proportions. Another well-established ML framework is PhyML,[56] which can internally optimize diverse evolutionary parameters.

Although similar to ML inference, Bayesian inference (BI) combines the prior probability of a phylogeny with the likelihood to produce a posterior probability

distribution of trees, which can be interpreted as the probability of those trees (or tree) being correct.[57] Clade support is estimated by summarizing this distribution of trees through consensus analysis. Bayesian phylogenies are estimated using Metropolis-coupled Markov chain Monte Carlo (MCMC) methods and both are implemented in programs such as MrBayes.[58] The output of the BI analysis must be evaluated to assure the MCMC chains have mixed well and converged; such tasks can be performed in Tracer.[59] Importantly, the best fitting model can vary across sites. For this reason, programs such as RAxML or MrBayes implement partition-specific (i.e., sites or genomic regions) models that can improve the accuracy of phylogenetic inferences.[60]

Often, gene trees differ even when sampled from the same population. This can be the result of molecular processes (e.g., recombination) or stochastic variation (e.g., lineage sorting). Whatever the case, one may want to check if individual gene topologies are significantly different since ignoring these processes may lead to biased parameter inferences.[61] Multiple ML topological tests have been developed for such purpose and several are implemented in CONSEL.[62]

New coalescent approaches have been developed to deal with stochastic variation in gene trees from multilocus molecular data and to estimate gene trees and species tree. Among such, BEST[63] and *BEAST[64] consider the effect of incomplete lineage sorting (ILS) by implementing the multispecies coalescent model into a Bayesian hierarchical model. When estimating evolutionary relationships among microbes using DNA sequences, the reticulating impact of recombination becomes a significant issue. If recombination is substantial, the evolutionary history of those sequences no longer fits a bifurcating model as those described before, and therefore a tree representation may fail to accurately portray a reasonable genealogy.[65] Under such circumstances, network approaches[66] can be used instead. Recently, Woolley et al.[67] have revised the most common algorithms for building phylogenetic networks and concluded that the union of maximum parsimonious (UMP) trees[68] performed the best. TCS[69] and SplitsTree[70] also performed well at estimating network gene genealogies. Finally, Didelot and Falush[71] have developed a Bayesian coalescent approach (ClonalFrame) that also takes homologous recombination into account while inferring clonal relationships between the members of a sample.

5.2.2 Population Dynamics

The evolution of DNA sequences in natural populations can be described with parameters such as recombination, mutation, growth, and selection rates. Indeed, the accurate estimation of these parameters is key for understanding the dynamics and evolutionary history of those populations, their epidemiology, the potential for and mode of evolution of antibiotic resistance, and ultimately for applying efficient public health control strategies. Population parameters are more efficiently estimated using explicit statistical models of evolution, such as the coalescent approach, hence here we describe some population parameter estimators based on such models.

Recombination is generally defined as the exchange of genetic information between two nucleotide sequences. It influences biological evolution at many different levels as well as affects the estimation of other parameters. Comprehensive assessment of statistical methods for detecting and estimating recombination rates were presented in

Martin et al.[72] and Posada et al.[73] These studies concluded that one should not rely on a single method to detect or estimate recombination. With this idea in mind, software packages such as RDP4[74] have been developed to implement a variety of methods for the same data set. RDP4 is a package that includes 12 recombination estimators and allows the user to draw conclusions based on the outcome of multiple tests. Another ML method to detect recombination is GARD,[75] which outperformed previously developed methods. In addition, programs such as LAMARC, LDhat, CodABC, and OmegaMap[76] (described in Pérez-Losada et al.[77]) can be used to estimate recombination rates and, therefore, to quantify the amount of observed recombination. Similarly, these methods can estimate genetic diversity, the most important population parameter. Reviews of classical and newer statistical methods for estimating genetic diversity have been published elsewhere.[78–81]

Another key parameter for characterizing microbial population dynamics is the growth rate, which reflects the variation of genetic diversity over time. Growth rates can be estimated under a certain demographic model (e.g., exponential) or without dependence on a prespecified model, for example, Skyride.[82] The latter approach is implemented in BEAST,[83] which also allows for the analysis of temporally spaced sequence data. Exponential growth rates and genetic diversity can also be estimated in LAMARC.

The standard method for estimating selection in protein-coding DNA sequences is through the nonsynonymous (d_N) to synonymous (d_S) amino acid substitution ratio d_N/d_S (ω). $\omega > 1$ indicates adaptive or diversifying selection, $\omega < 1$ purifying selection, and $\omega \approx 1$ lack of selection (neutral evolution). ω is usually estimated within an ML phylogenetic framework and assuming an explicit model of codon substitution. Such models can be very complex, allowing, for example, ω to vary across amino acid sites and/or tree branches, for example, Yang.[84] If significant evidence (usually obtained through likelihood ratio tests, LRT) of adaptive selection is obtained, then Bayesian tests can be applied to detect amino acid sites under selection, for example, Yang et al.[85] Such methods are implemented and described in more detail in the software package PAML.[84] However, if recombination is suspected in the data, it should be considered when estimating ω to avoid false positively selected sites.[61] Thus, one could estimate recombination and selection rates simultaneously with OmegaMap or CodABC, or account for the former while estimating the latter, for example, HYPHY.[86]

Other key factors in microbial dynamics are time of emergence (e.g., pathogen outbreaks) and geographic distribution of pathogens. New probabilistic models were developed within the Bayesian framework[87] for inference and hypothesis testing of divergence times, ancestral locations and historical patterns of migration (i.e., phylogeographical history). Such models are implemented in BEAST and SPREAD[88] and visualized using virtual globe software, such as Google Earth; they have already begun to be applied to the analysis of MLST and/or genome and SNP data.[89,90]

Most of the nucleotide-based methods described earlier, and others, have been implemented in user-friendly web servers, such as CBSU (cbsuapps.tc.cornell.edu), CIPRES (www.phylo.org), Datamonkey (www.datamonkey.org), or PhyML (www.atgc-montpellier.fr/phyml/).

6. Applications of Multilocus Sequence Typing

MLST analysis and databases are standardized and broadly used, filled with historical information, and firmly established in molecular and clinical laboratories worldwide. Consequently, new typing applications are seeking to integrate existing MLST schemes with whole-genome shotgun data to characterize microbial populations, rather than creating from scratch new typing methods. MLST is probably the most flexible typing method since it can be implemented in small laboratories with standard equipment (PCR + Sanger sequencing), as well as in medium-sized facilities with vanguard infrastructure (targeted-amplicon sequencing; AmpliSeq panels, robotics, and so on) or laboratories with whole-genome sequencing capability (in silico MLST).

Although primarily developed for the characterization of organisms (typing), MLST sequence data have also been applied to other endeavors such as molecular epidemiology (e.g., disease transmission and surveillance programs) and public health (e.g., monitor and evaluate vaccination programs), as well as to other areas such as phylogenetics, taxonomy, speciation, population genetics, biosafety, and even to the inference of human migrations.

6.1 Molecular Epidemiology and Public Health

MLST has gained widespread popularity as a typing method and its use has advanced understanding of bacterial evolution and has provided insights into the epidemiology of bacterial diseases. In the context of surveillance and management of disease outbreaks, being able to quickly type and track infectious diseases is of paramount importance. Many studies exemplify the use of MLST in these circumstances: emergence of zoonosis,[89,90] detection of disease outbreaks,[91,92] estimation of prevalence rates,[93,94] and the origins of virulence factors (vertical or horizontally transmitted).[95,96]

MLST data have been also used to infer population structure and study the emergence and spread of antibiotic resistance.[97] For example, MLST has been used to diagnose human-associated population structure in the opportunistic pathogen *Ochrobactrum anthropi*. Romano et al.[98] developed an MLST scheme for this pathogen and used the evolutionary information inherent in the DNA sequences to identify a human-associated subpopulation from their collection of clinical and environmental isolates. Likewise, MLST has been used to track drug-resistance variants through patients. Oteo et al.[99] collected 162 isolates of *Klebsiella pneumoniae* from five hospitals in Spain and used the MLST data to demonstrate the spreading of this bacteria as pathogen and colonizer of newborns and adult patients with multilocus resistance acquired through recombination. Similarly, Lee et al.[100] used MLST to identify epidemic and virulent ciprofloxacin-resistant *E. coli* clones and their population structure in Korea causing urinary tract infections.

In a number of studies, MLST data have been used to reveal the epidemiological history of infectious diseases. For example, MLST has been successful in identifying clinically important strains of *Neisseria meningitides*, that is, hyperinvasive lineages.[101] MLST has been applied to a number of clinically important bacterial populations, including hospital-acquired strains of *Enterococcus faecalis* and

Enterococcus faecium,[102,103] and *Streptococcus pneumoniae* strains associated with invasive disease.[104] In some cases, MLST has failed to distinguish clinically relevant populations. For example, *Staphylococcus aureus* isolates from persons with nasal carriage, community-acquired pneumonia, and hospital-acquired invasive disease are evenly distributed among clonal complexes.[105] Similarly, there is a poor correlation between MLST data and tissue tropisms (throat or skin) of *Streptococcus pyogenes* isolates.[106] For phenotypes that are based on one or a few genes, such as antibiotic resistance, correlations with MLST data have been large. The evolutionary history of methicillin-resistant *S. aureus* (MRSA) has been clarified by MLST data, including the typing of the methicillin-resistance genetic element, SCCmec.[107] Along the same lines, MLST has been used to identify transmission chains as demonstrated by Choudhury et al.[108] where the authors identified outbreak sources and characterized outbreaks of gonorrhea. They typed consecutive *gonococcal* strains from London STI clinics over a 9-month period. Clusters of patients with the same strain showed similarities in behavioral and demographic features, suggesting that different strain clusters represent localized transmission chains.

New phylogenetic coalescent models have been developed allowing researchers to infer from genetic data more familiar parameters, such as the reproductive number of viruses,[109] as well as to model epidemiological dynamics that describe changes in population size or date of origin.[110–113] Lastly, examples of MLST and whole-genome sequencing integration abound (see Pérez-Losada et al.[5]). In molecular epidemiology, studies since 2010 combine MLST data with in silico MLST in an effort to put new isolates in context without losing the resolution and insight gained by having the full genetic complement of the bacteria in question.[114–117]

6.2 Species Diagnosis and Phylogenetics

MLST data have been used to distinguish similar species, to inform the division of a genus into species, and to ask whether bacterial species exist. The MLST data are especially useful for species diagnoses as they provide both genealogical information as well as information on recombination.[118] Indeed, even when the MLST are not as discriminating as other approaches, the phylogenetic information available through MLST provides novel insights into species and strain relatedness that impact public health decisions. In a study of *Clostridium difficile*, for example, Marsh et al.[28] found MLST less discriminatory compared to MLVA or restriction endonuclease analysis (REA) although concordant, but the combination of MLST with MLVA provided novel insights into the origins and evolutionary relationships bearing clinical and public health importance. Similarly, a phylogenetic analysis of concatenated sequences of seven MLST loci for *Bacillus pseudomallei* and *Bacillus thailandensis*, both soil saprophytes, and *Bacillus mallei*, the cause of glanders, showed that all *B. pseudomallei* strains were tightly clustered and well resolved from all *B. thailandensis* strains.[119] However, *B. mallei* clustered with *B. pseudomallei* and, although designated as a "species," can be considered to be a strain (or clone) of *B. pseudomallei*. Other examples of bacterial species that are actually clones with distinctive biology and ecology include *Bacillus anthracis*[120] and *Salmonella typhi*.[121] *Neisseria gonorrhoeae* strains form a

tight cluster at the end of a long branch arising from the meningococcal cluster,[122] supporting the hypothesis that gonococci arose relatively recently as a strain of human pharyngeal *Neisseria* species that acquired the ability to colonize the genital tract and be transmitted by the sexual route.[123] MLST has also proven useful in the context of taxonomic groups with low genomic representation, for example, neglected diseases or industrial microbes,[124,125] and on studies where large numbers of samples are analyzed.[126,127] For instance, Nuñez et al. interrogated the genetic structure of the bioleaching microbe *Acidithiobacillus caldus* and found overall low genetic diversity from different geographic locations, which supports current taxonomic assignments and suggests that bioprocesses constrain genetic diversity.[125]

7. Conclusions and Prospects

MLST has become a standard and flexible approach for characterizing bacteria and some eukaryotes mainly due to the existence of comprehensive databases and its broad implementation in clinical laboratory settings, from basic research laboratories (PCR + Sanger) to core sequencing facilities (cgMLST; in silico MLST). MLST has expanded its basic scheme to incorporate more and new molecular markers, such as ribosomal proteins and large matrices of orthologous genes (gene-by-gene approach), and more recently, to integrate pan and core-genome concepts as well as draft and full genomes. Two-tier strategies currently being applied to human microbiome research where investigations start by using MLST to type as many samples as possible, and continue by delving further into isolate groups of particular interest by using whole-genome sequencing are already in practice.[117,128]

New MLST-genome strategies will also provide more accurate and robust estimates of population genetic parameters under more complex and realistic statistical models such as those based on the coalescent model.[129] Moreover, within this framework, epidemiological data can also be integrated; hence more comprehensive and faster assessments of pathogen dynamics can be achieved. Microbial genomics is expanding outside research laboratories into clinical practice and molecular diagnostics.[130,131] One can only assume that classical or expanded forms of MLST will remain a key component of the microbial genomicist's toolkit toward understanding the ecology and evolution of infectious diseases.

Acknowledgments

M.P.-L. was funded by a DC D-CFAR Research Award from the District of Columbia Developmental Center for AIDS Research (P30AI087714), by a University Facilitating Fund award from George Washington University, and a K12 Career Development Program 5 K12 HL119994 award. M.A. was supported by the Portuguese Government through the FCT Starting Grant IF/00955/2014. E.C.N. was funded by "CONICYT + PAI/CONCURSO NACIONAL APOYO AL RETORNO DE INVESTIGADORES/AS DESDE EL EXTRANJERO, CONVOCATORIA 2014 + FOLIO 82140008."

References

1. Cooper JE, Feil EJ. Multilocus sequence typing—what is resolved? *Trends Microbiol* 2004;**12**:373–7.
2. Foley SL, Lynne AM, Nayak R. Molecular typing methodologies for microbial source tracking and epidemiological investigations of gram-negative bacterial foodborne pathogens. *Infect Genet Evol* 2009;**9**:430–40.
3. Maiden MC, Bygraves JA, Feil E, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci USA* 1998;**95**:3140–5.
4. Maiden MC. Multilocus sequence typing of bacteria. *Annu Rev Microbiol* 2006;**60**: 561–88.
5. Pérez-Losada M, Cabezas P, Castro-Nallar E, Crandall KA. Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. *Infect Genet Evol* 2013; **16**:38–53.
6. Jolley KA, Maiden MC. Using MLST to study bacterial variation: prospects in the genomic era. *Future Microbiol* 2014;**9**:623–30.
7. Maiden MC, van Rensburg MJJ, Bray JE, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat Rev Microbiol* 2013;**11**:728–36.
8. Urwin R, Maiden MC. Multi-locus sequence typing: a tool for global epidemiology. *Trends Microbiol* 2003;**11**:479–87.
9. Sullivan CB, Diggle MA, Clarke SC. Multilocus sequence typing: data analysis in clinical microbiology and public health. *Mol Biotechnol* 2005;**29**:245–54.
10. Boers SA, van der Reijden WA, Jansen R. High-throughput multilocus sequence typing: bringing molecular typing to the next level. *PLoS One* 2012;**7**:e39630.
11. Jolley KA, Bliss CM, Bennett JS, et al. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology* 2012;**158**: 1005–15.
12. Larsen MV, Cosentino S, Rasmussen S, et al. Multilocus sequence typing of total-genome-sequenced bacteria. *J Clin Microbiol* 2012;**50**:1355–61.
13. Pérez-Losada M, Browne EB, Madsen A, Wirth T, Viscidi RP, Crandall KA. Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. *Infect Genet Evol* 2006;**6**:97–112.
14. Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005;**437**:376–80.
15. Chen Y, Frazzitta AE, Litvintseva AP, et al. Next generation multilocus sequence typing (NGMLST) and the analytical software program MLSTEZ enable efficient, cost-effective, high-throughput, multilocus sequencing typing. *Fungal Genet Biol* 2015;**75**: 64–71.
16. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol* 2013;**79**:5112–20.
17. O'Halloran DM. PrimerMapper: high throughput primer design and graphical assembly for PCR and SNP detection. *Sci Rep* 2016;**6**:20631.
18. Untergasser A, Cutcutache I, Koressaar T, et al. Primer3—new capabilities and interfaces. *Nucleic Acids Res* 2012;**40**:e115.
19. Pérez-Losada M, Crandall KA, Zenilman J, Viscidi RP. Temporal trends in gonococcal population genetics in a high prevalence urban community. *Infect Genet Evol* 2007;**7**: 271–8.

20. Carattoli A, Zankari E, García-Fernández A, et al. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother* 2014;**58**:3895−903.
21. Yoshida C, Kruczkiewicz P, Laing C, et al. The *Salmonella In Silico* Typing Resource (SISTR): an open web-accessible tool for rapidly typing and subtyping draft *Salmonella* genome assemblies. *PLoS One* 2015;**11**:e0147101.
22. Kruczkiewicz P, Mutschall S, Barker D, et al. MIST: a tool for rapid in silico generation of molecular data from bacterial genome sequences. *Bioinformatics* 2013:316−23.
23. Inouye M, Dashnow H, Raven L-A, et al. SRST2: rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med* 2014;**6**:1−16.
24. Cookson BD, Robinson DA, Monk AB, et al. Evaluation of molecular typing methods in characterizing a European collection of epidemic methicillin-resistant *Staphylococcus aureus* strains: the HARMONY collection. *J Clin Microbiol* 2007;**45**:1830−7.
25. Metzgar D, Baynes D, Hansen CJ, et al. Inference of antibiotic resistance and virulence among diverse group A *Streptococcus* strains using *emm* sequencing and multilocus genotyping methods. *PLoS One* 2009;**4**:e6897.
26. Siarkou VI, Vorimore F, Vicari N, et al. Diversification and distribution of Ruminant *Chlamydia abortus* clones assessed by mlst and MLVA. *PLoS One* 2015;**10**:e0126433.
27. van Cuyck H, Pichon B, Leroy P, et al. Multiple-locus variable-number tandem-repeat analysis of *Streptococcus pneumoniae* and comparison with multiple loci sequence typing. *BMC Microbiol* 2012;**12**:241.
28. Marsh JW, O'Leary MM, Shutt KA, et al. Multilocus variable-number tandem-repeat analysis and multilocus sequence typing reveal genetic relationships among *Clostridium difficile* isolates genotyped by restriction endonuclease analysis. *J Clin Microbiol* 2010;**48**: 412−8.
29. de Been M, Pinholt M, Top J, et al. A core genome MLST scheme for high-resolution typing of *Enterococcus faecium*. *J Clin Microbiol* 2015.
30. Hall BG, Ehrlich GD, Hu FZ. Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. *Microbiology* 2010;**156**:1060−8.
31. Leekitcharoenphon P, Lukjancenko O, Friis C, Aarestrup FM, Ussery DW. Genomic variation in *Salmonella enterica* core genes for epidemiological typing. *BMC Genomics* 2012;**13**:88.
32. Chan MS, Maiden MC, Spratt BG. Database-driven multi locus sequence typing (MLST) of bacterial pathogens. *Bioinformatics* 2001;**17**:1077−83.
33. Jolley KA, Chan MS, Maiden MC. mlstdbNet − distributed multi-locus sequence typing (MLST) databases. *BMC Bioinformatics* 2004;**5**:86.
34. Jolley KA, Maiden MC. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 2010;**11**:595.
35. Jolley KA, Maiden MC. AgdbNet − antigen sequence database software for bacterial typing. *BMC Bioinformatics* 2006;**7**:314.
36. Jefferies J, Clarke SC, Diggle MA, Smith A, Dowson C, Mitchell T. Automated pneumococcal MLST using liquid-handling robotics and a capillary DNA sequencer. *Mol Biotechnol* 2003;**24**:303−7.
37. Parkhill J, Sebaihia M, Preston A, et al. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nat Genet* 2003;**35**:32−40.
38. Enright MC, Spratt BG. Multilocus sequence typing. *Trends Microbiol* 1999;**7**:482−7.
39. Loubna T, Pérez-Losada M, Gu W, et al. Population dynamics of *Neisseria gonorrhoeae* in Shanghai, China: a comparative study. *BMC Infect Dis* 2010;**10**:13.

40. Jolley KA, Feil EJ, Chan MS, Maiden MC. Sequence type analysis and recombinational tests (START). *Bioinformatics* 2001;**17**:1230−1.
41. Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J Bacteriol* 2004;**186**:1518−30.
42. Francisco AP, Bugalho M, Ramirez M, Carrico JA. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. *BMC Bioinformatics* 2009;**10**: 152.
43. Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M, Carrio JA. PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. *BMC Bioinformatics* 2012;**13**:87.
44. Wiens JJ. Reconstructing phylogenies from allozyme data: comparing method performance with congruence. *Biol J Linn Soc* 2000;**70**:613−32.
45. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013;**30**:772−80.
46. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;**32**:1792−7.
47. Lemmon AR, Moriarty EC. The importance of proper model assumption in bayesian phylogenetics. *Syst Biol* 2004;**53**:265−77.
48. Posada D, Buckley TR. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and Bayesian approaches over likelihood ratio tests. *Syst Biol* 2004;**53**:793−808.
49. Arenas M. Trends in substitution models of molecular evolution. *Front Genet* 2015;**6**:319.
50. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* 2012;**9**:772.
51. Luo A, Qiao H, Zhang Y, et al. Performance of criteria for selecting evolutionary models in phylogenetics: a comprehensive study based on simulated datasets. *BMC Evol Biol* 2010; **10**:242.
52. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 1981;**17**:368−76.
53. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 2006;**22**:2688−90.
54. Stamatakis A, Aberer AJ, Goll C, Smith SA, Berger SA, Izquierdo-Carrasco F. RAxML-Light: a tool for computing terabyte phylogenies. *Bioinformatics* 2012;**28**:2064−6.
55. Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 1985;**39**:783−91.
56. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 2010;**59**:307−21.
57. Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 2001;**294**:2310−4.
58. Ronquist F, Teslenko M, van der Mark P, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* 2012;**61**:539−42.
59. Rambaut A, Drummond AJ. *Tracer: MCMC trace analysis tool*. 1.5 ed. Edinburgh: Institute of Evolutionary Biology; 2009. http://tree.bio.ed.ac.uk/software/tracer/.
60. Zoller S, Boskova V, Anisimova M. Maximum-likelihood tree estimation using codon substitution models with multiple partitions. *Mol Biol Evol* 2015;**32**:2208−16.
61. Arenas M, Posada D. Coalescent simulation of intracodon recombination. *Genetics* 2010; **184**:429−37.

62. Shimodaira H, Hasegawa M. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 2001;**17**:1246−7.
63. Liu L. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics* 2008;**24**:2542−3.
64. Heled J, Drummond AJ. Bayesian inference of species trees from multilocus data. *Mol Biol Evol* 2010;**27**:570−80.
65. Schierup MH, Hein J. Consequences of recombination on traditional phylogenetic analysis. *Genetics* 2000;**156**:879−91.
66. Posada D, Crandall KA. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci USA* 2001;**98**:13757−62.
67. Woolley SW, Posada D, Crandall KA. A comparison of phylogenetic network methods using computer simulation. *PLoS Comput Biol* 2008;**3**:e1913.
68. Cassens I, Mardulyn P, Milinkovitch MC. Evaluating intraspecific "network" construction methods using simulated sequence data: do existing algorithms outperform the global maximum parsimony approach? *Syst Biol* 2005;**54**:363−72.
69. Templeton AR, Crandall KA, Sing CF. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* 1992;**132**:619−33.
70. Huson DH. SplitsTree: a program for analyzing and visualizing evolutionary data. *Bioinformatics* 1998;**14**:68−73.
71. Didelot X, Falush D. Inference of bacterial microevolution using multilocus sequence data. *Genetics* 2007;**175**:1251−66.
72. Martin DP, Lemey P, Posada D. Analysing recombination in nucleotide sequences. *Mol Ecol Resour* 2011;**11**:943−55.
73. Posada D, Crandall KA, Holmes EC. Recombination in evolutionary genomics. *Annu Rev Genet* 2002;**36**:75−97.
74. Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* 2010;**26**:2462−3.
75. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. GARD: a genetic algorithm for recombination detection. *Bioinformatics* 2006;**22**:3096−8.
76. Arenas M, Lopes JS, Beaumont MA, Posada D. CodABC: a computational framework to coestimate recombination, substitution, and molecular adaptation rates by approximate Bayesian computation. *Mol Biol Evol* 2015;**32**:1109−12.
77. Pérez-Losada M, Porter ML, Tazi L, Crandall KA. New methods for inferring population dynamics from microbial sequences. *Infect Genet Evol* 2007;**7**:24−43.
78. Pearse DE, Crandall K. Beyond F_{ST}: analysis of population genetic data for conservation. *Conserv Genet* 2004;**5**:585−602.
79. Excoffier L, Heckel G. Computer programs for population genetics data analysis: a survival guide. *Nat Rev Genet* 2006;**7**:745−58.
80. Waples RS, Gaggiotti O. 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* 2006;**15**:1419−39.
81. Bashalkhanov S, Pandey M, Rajora OP. A simple method for estimating genetic diversity in large populations from finite sample sizes. *BMC Genet* 2009;**10**:84.
82. Minin VN, Bloomquist EW, Suchard MA. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol Biol Evol* 2008;**25**: 1459−71.
83. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 2012;**29**:1969−73.

84. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 2007;**24**: 1586−91.
85. Yang Z, Wong WS, Nielsen R. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol* 2005;**22**:1107−18.
86. Kosakovsky Pond SL, Frost SDW, Muse SV. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 2005;**21**:676−9.
87. Lemey P, Rambaut A, Welch JJ, Suchard MA. Phylogeography takes a relaxed random walk in continuous space and time. *Mol Biol Evol* 2010;**27**:1877−85.
88. Bielejec F, Rambaut A, Suchard MA, Lemey P. SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics. *Bioinformatics* 2011;**27**:2910−2.
89. McAdam PR, Templeton KE, Edwards GF, et al. Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant *Staphylococcus aureus*. *Proc Natl Acad Sci USA* 2012;**109**:9107−12.
90. Weinert LA, Welch JJ, Suchard MA, Lemey P, Rambaut A, Fitzgerald JR. Molecular dating of human-to-bovid host jumps by *Staphylococcus aureus* reveals an association with the spread of domestication. *Biol Lett* 2012;**8**:829−32.
91. Palazzo IC, Pitondo-Silva A, Levy CE, da Costa Darini AL. Changes in vancomycin-resistant *Enterococcus faecium* causing outbreaks in Brazil. *J Hosp Infect* 2011;**79**:70−4.
92. Vanderkooi OG, Church DL, MacDonald J, Zucol F, Kellner JD. Community-based outbreaks in vulnerable populations of invasive infections caused by *Streptococcus pneumoniae* serotypes 5 and 8 in Calgary, Canada. *PLoS One* 2011;**6**:e28547.
93. Haran KP, Godden SM, Boxrud D, Jawahir S, Bender JB, Sreevatsan S. Prevalence and characterization of *Staphylococcus aureus*, including methicillin-resistant *Staphylococcus aureus*, isolated from bulk tank milk from Minnesota dairy farms. *J Clin Microbiol* 2012; **50**:688−95.
94. Ibarz-Pavon AB, Morais L, Sigauque B, et al. Epidemiology, molecular characterization and antibiotic resistance of *Neisseria meningitidis* from patients ≤15 years in Manhica, rural Mozambique. *PLoS One* 2011;**6**:e19717.
95. Martin V, Maldonado-Barragan A, Moles L, et al. Sharing of bacterial strains between breast milk and infant feces. *J Hum Lact* 2012;**28**:36−44.
96. Walker AS, Eyre DW, Wyllie DH, et al. Characterisation of *Clostridium difficile* hospital ward-based transmission using extensive epidemiological data and molecular typing. *PLoS Med* 2012;**9**:e1001172.
97. Egger R, Korczak BM, Niederer L, Overesch G, Kuhnert P. Genotypes and antibiotic resistance of *Campylobacter coli* in fattening pigs. *Vet Microbiol* 2012;**155**:272−8.
98. Romano S, Aujoulat F, Jumas-Bilak E, et al. Multilocus sequence typing supports the hypothesis that *Ochrobactrum anthropi* displays a human-associated subpopulation. *BMC Microbiol* 2009;**9**.
99. Oteo J, Cuevas O, Lopez-Rodriguez I, et al. Emergence of CTX-M-15-producing *Klebsiella pneumoniae* of multilocus sequence types 1, 11, 14, 17, 20, 35 and 36 as pathogens and colonizers in newborns and adults. *J Antimicrob Chemother* 2009;**64**:524−8.
100. Lee MY, Choi HJ, Choi JY, et al. Dissemination of ST131 and ST393 community-onset, ciprofloxacin-resistant *Escherichia coli* clones causing urinary tract infections in Korea. *J Infect* 2010;**60**:146−53.
101. Yazdankhah SP, Kriz P, Tzanakaki G, et al. Distribution of serogroups and genotypes among disease-associated and carried isolates of *Neisseria meningitidis* from the Czech Republic, Greece, and Norway. *J Clin Microbiol* 2004;**42**:5146−53.

102. Ruiz-Garbajosa P, Bonten MJ, Robinson DA, et al. Multilocus sequence typing scheme for *Enterococcus faecalis* reveals hospital-adapted genetic complexes in a background of high rates of recombination. *J Clin Microbiol* 2006;**44**:2220−8.
103. Leavis HL, Bonten MJ, Willems RJ. Identification of high-risk enterococcal clonal complexes: global dispersion and antibiotic resistance. *Curr Opin Microbiol* 2006;**9**: 454−60.
104. Enright MC, Spratt BG. A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. *Microbiology* 1998;**144**(Pt 11):3049−60.
105. Feil EJ, Cooper JE, Grundmann H, et al. How clonal is *Staphylococcus aureus*? *J Bacteriol* 2003;**185**:3307−16.
106. Kalia A, Spratt BG, Enright MC, Bessen DE. Influence of recombination and niche separation on the population genetic structure of the pathogen *Streptococcus pyogenes*. *Infect Immun* 2002;**70**:1971−83.
107. Robinson DA, Enright MC. Evolutionary models of the emergence of methicillin-resistant *Staphylococcus aureus*. *Antimicrob Agents Chemother* 2003;**47**:3926−34.
108. Choudhury B, Risley CL, Ghani AC, et al. Identification of individuals with gonorrhoea within sexual networks: a population-based study. *Lancet* 2006;**368**:139−46.
109. Stadler T, Kouyos R, von Wyl V, et al. Estimating the basic reproductive number from viral sequence data. *Mol Biol Evol* 2012;**29**:347−57.
110. Popinga A, Vaughan T, Stadler T, Drummond AJ. Inferring epidemiological dynamics with Bayesian coalescent inference: the merits of deterministic and stochastic models. *Genetics* 2015;**199**:595−607.
111. du Plessis L, Stadler T. Getting to the root of epidemic spread with phylodynamic analysis of genomic data. *Trends Microbiol* 2015;**23**:383−6.
112. Volz EM, Frost SD. Inferring the source of transmission with phylogenetic data. *PLoS Comput Biol* 2013;**9**:e1003397.
113. Rasmussen DA, Volz EM, Koelle K. Phylodynamic inference for structured epidemiological models. *PLoS Comput Biol* 2014;**10**:e1003570.
114. Hamby SE, Joseph S, Forsythe SJ, Chuzhanova N. In silico identification of pathogenic strains of *Cronobacter* from biochemical data reveals association of inositol fermentation with pathogenicity. *BMC Microbiol* 2011;**11**:1.
115. Stasiewicz MJ, Oliver HF, Wiedmann M, den Bakker HC. Whole-genome sequencing allows for improved identification of persistent *Listeria monocytogenes* in food-associated environments. *Appl Environ Microbiol* 2015;**81**:6024−37.
116. Wong VK, Baker S, Pickard DJ, et al. Phylogeographical analysis of the dominant multidrug-resistant H58 clade of *Salmonella* Typhi identifies inter-and intracontinental transmission events. *Nat Genet* 2015;**47**:632−9.
117. Holt KE, Wertheim H, Zadoks RN, et al. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. *Proc Natl Acad Sci* 2015;**112**:E3574−81.
118. Fraser C, Hanage WP, Spratt BG. Recombination and the nature of bacterial speciation. *Science* 2007;**315**:476−80.
119. Godoy D, Randle G, Simpson AJ, et al. Multilocus sequence typing and evolutionary relationships among the causative agents of melioidosis and glanders, *Burkholderia pseudomallei* and *Burkholderia mallei*. *J Clin Microbiol* 2003;**41**:2068−79.
120. Priest FG, Barker M, Baillie LW, Holmes EC, Maiden MC. Population structure and evolution of the *Bacillus cereus* group. *J Bacteriol* 2004;**186**:7959−70.

121. Kidgell C, Reichard U, Wain J, et al. *Salmonella typhi*, the causative agent of typhoid fever, is approximately 50,000 years old. *Infect Genet Evol* 2002;**2**:39−45.
122. Hanage WP, Fraser C, Spratt BG. Sequences, sequence clusters and bacterial species. *Philos Trans R Soc Lond B Biol Sci* 2006;**361**:1917−27.
123. Vazquez JA, de la Fuente L, Berron S, et al. Ecological separation and genetic isolation of *Neisseria gonorrhoeae* and *Neisseria meningitidis*. *Curr Biol* 1993;**3**:567−72.
124. Boonsilp S, Thaipadungpanit J, Amornchai P, et al. A single multilocus sequence typing (MLST) scheme for seven pathogenic *Leptospira* species. *PLoS Negl Trop Dis* 2013;**7**: e1954.
125. Nuñez H, Loyola D, Cárdenas JP, Holmes DS, Johnson DB, Quatrini R. Multilocus sequence typing scheme for *Acidithiobacillus caldus* strain evaluation and differentiation. *Res Microbiol* 2014;**165**:735−42.
126. Jacquot M, Bisseux M, Abrial D, et al. High-throughput sequence typing reveals genetic differentiation and host specialization among populations of the *Borrelia burgdorferi* species complex that infect rodents. *PLoS One* 2014;**9**:e88581.
127. Rosales R, Churchward C, Schnee C, et al. Global multilocus sequence typing analysis of *Mycoplasma bovis* isolates reveals two main population clusters. *J Clin Microbiol* 2015; **53**:789−94.
128. Franzosa EA, Hsu T, Sirota-Madi A, et al. Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. *Nat Rev Microbiol* 2015;**13**:360−72.
129. Mather AE, Vaughan TG, French NP. Molecular approaches to understanding transmission and source attribution in nontyphoidal *Salmonella* and their application in Africa. *Clin Infect Dis* 2015;**61**:S259−65.
130. Gardy J, Loman NJ, Rambaut A. Real-time digital pathogen surveillance—the time is now. *Genome Biol* 2015;**16**.
131. Luheshi LM, Raza S, Peacock SJ. Moving pathogen genomics out of the lab and into the clinic: what will it take? *Genome Med* 2015;**7**.