Infection, Genetics and Evolution 16 (2013) 38–53

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Review

Pathogen typing in the genomics era: MLST and the future of molecular epidemiology

Marcos Pérez-Losada [a,*], Patricia Cabezas [b,c,d], Eduardo Castro-Nallar [b,c], Keith A. Crandall [b,c]

[a] CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, Campus Agrário de Vairão, 4485-661 Vairão, Portugal
[b] Department of Biological Sciences, George Washington University, Washington, DC 20052, USA
[c] Computational Biology Institute, George Washington University, Ashburn, VA 20147, USA
[d] Department of Biology, Brigham Young University, Provo, UT 84602 USA

ARTICLE INFO

Article history:
Received 25 September 2012
Received in revised form 11 January 2013
Accepted 15 January 2013
Available online 26 January 2013

Keywords:
Multi-locus sequence typing (MLST)
Next-generation sequence data
Molecular epidemiology
Pathogen evolution

ABSTRACT

Multi-locus sequence typing (MLST) is a high-resolution genetic typing approach to identify species and strains of pathogens impacting human health, agriculture (animals and plants), and biosafety. In this review, we outline the general concepts behind MLST, molecular approaches for obtaining MLST data, analytical approaches for MLST data, and the contributions MLST studies have made in a wide variety of areas. We then look at the future of MLST and their relative strengths and weaknesses with respect to whole genome sequence typing approaches that are moving into the research arena at an ever-increasing pace. Throughout the paper, we provide exemplar references of these various aspects of MLST. The literature is simply too vast to make this review comprehensive, nevertheless, we have attempted to include enough references in a variety of key areas to introduce the reader to the broad applications and complications of MLST data.

Contents

[*] Corresponding author. Tel.: +351 252 660411; fax: +351 252 661780.
E-mail address: mlosada323@gmail.com (M. Pérez-Losada).

1567-1348/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.meegid.2013.01.009

1. Introduction

The vast majority of bacteria are harmless or beneficial, but the few pathogenic strains are a major cause of human disease and death. Bacterial pathogens are the etiological agents of a wide range of infections including syphilis, cholera and tuberculosis among others. Understanding the processes controlling transmission relies first and foremost on the ability to identify and accurately distinguish between strains of infectious pathogens. Accurate and efficient strain identification is also essential for epidemiological surveillance and subsequent design of public health control strategies (Comas et al., 2009; Schulte and Perera, 1993). Over the last decades, different molecular techniques have been extensively exploited to identify isolates and localize disease outbreaks, but their poor portability usually hindered, rather than elucidated, pathogen epidemiology (Maiden, 2006; Urwin and Maiden, 2003). To overcome this problem, molecular microbiology took advantage of existing knowledge on bacterial evolution and population biology, easy access and low cost of high-throughput Sanger sequencing, and internet databasing resources, to propose the nucleotide sequence-based approach of multilocus sequence typing (MLST; Maiden et al., 1998). This procedure allows for the unambiguous characterization of isolates from infectious agents using sequences of internal fragments of usually seven housekeeping genes (i.e., constitutive genes required for the maintenance of basic cellular functions). Gene regions of approximately 450–500 bp are sequenced and those found unique within a species are assigned an allele number. Each isolate is then characterized by the alleles at each of the seven loci, which constitute its allelic profile or sequence type (ST).

The MLST approach provides an accurate assessment of species and sometimes even strains and has the added advantage of also providing population genetic insights into levels and directionality of gene flow. This genetic based species diagnosis is much more accurate than performing conventional immunological assays to determine species and strain. Often, these phenotypic assays do not reflect underlying genealogical information (e.g., Lewis-Rogers et al., 2009). Thus, misdiagnoses can easily occur without relevant genealogical information analyzed in an evolutionary and population genetic framework (Crandall and Pérez-Losada, 2008). MLST approaches provide such high-resolution genealogical data.

The first studies on bacterial population structure in the 1980's were fundamental to the development of MLST (Feil et al., 1999). These studies revealed genetic exchange through recombination as a major driving force in the evolution of most prokaryotes (Maiden, 2006). This finding changed the predominant paradigm of the "clonal model" in bacterial population genetics to a broader concept of panmictic and partially clonal models (Smith et al., 1993). Consequently, inferring genetic relatedness among isolates based on single markers was unreliable and a new method was needed that compared information from across multiple independent markers. The MLST scheme played a major role in investigating the extent of genetic structure in bacterial populations and rapidly became the cornerstone technique for molecular typing of pathogenic microorganisms (Maiden, 2006).

As currently used, MLST has achieved high levels of discrimination and has provided meaningful data to understand the evolution and epidemiology of pathogens. But given the recent advances in sequencing technologies, the question naturally arises: what is the future of the MLST scheme in the genomic era? Technological advances in high-throughput genome sequencing platforms (e.g., 454 Roche, Illumina/Solexa, Ion Torrent, and ABI SOLiD) glimpse a promising scenario to improve the resolution of molecular epidemiological studies to the most accurate level ever seen and will likely provide unprecedented insights into the evolution of bacterial populations. Here we review the past, present, and future of the MLST approach. Because of the extensive literature published on the topic, this review cannot be comprehensive in its scope. Instead, we provide a summary on how the MLST scheme transformed molecular epidemiological studies (Section 1), it is now integrated within the next-generation sequencing techniques (Section 2), it can be efficiently analyzed (Section 3) and contributed to understand molecular epidemiology and evolution of bacterial pathogens (Section 4). Moreover, we discuss the future of MLST approaches in the genomic era as whole genome data are rapidly becoming available for pathogen studies (Section 5).

1.1. MLST databases: origins and recent advances via internet resources

The MLST approach provided for the first time the reproducibility and portability needed to develop a worldwide pathogen-typing database easily accessible to public health and research communities. The MLST scheme was first developed and available via the Internet for the species *Neisseria meningitidis* (Maiden et al., 1998), and this trend grew rapidly to include other bacterial species (Enright and Spratt, 1998; Heym et al., 2002; Kriz et al., 2002). The first MLST website was implemented early on in the software MLSTdB, which was structured as a single combined database (Chan et al., 2001). This first online resource worked well for the small datasets initially produced, but as the number of schemes available increased, several limitations as data redundancy, isolate bias and access became apparent (Pérez-Losada et al., 2011). Consequently, a reworked version of the original software, namely MLSTdbNET (Jolley et al., 2004), was developed in order to provide a network database structure. The premise behind this new tool was the creation of separated databases to store isolate-specific information and allelic profiles, so that any number of isolate databases could be constructed. Those databases are actively curated to avoid the accumulation of sequencing errors that could lead to illusory alleles and ST profiles (Jolley, 2009). However, data retrieved from the databases comprise reported diversity, but are unstructured and do not necessarily represent natural populations (Urwin and Maiden, 2003).

MLST databases are now available for at least 79 organisms (75 for bacteria, 3 for fungi and 1 protozoan) and offer three main types of queries: (1) allele sequence identification and comparison, (2) allelic profile identification and comparison and (3) matching of isolates. Most MLST schemes are available at the websites hosted at the University of Oxford in the United Kingdom (http://pubmlst.org) and the United Kingdom's Imperial College (http://www.mlst.net), although some schemes can also be found at the Environmental Research Institute, Cork, Ireland (http://mlst.ucc.ie) and the Pasteur Institute, Paris (http://www.pasteur.fr/mlst). The international mirrored PubMLST website provides access to the abovementioned MLSTdbNET database, but also to the antigen sequence software (agdbNET) for bacterial typing (Jolley and Maiden, 2006), and to the recently developed Bacterial Isolate Genome

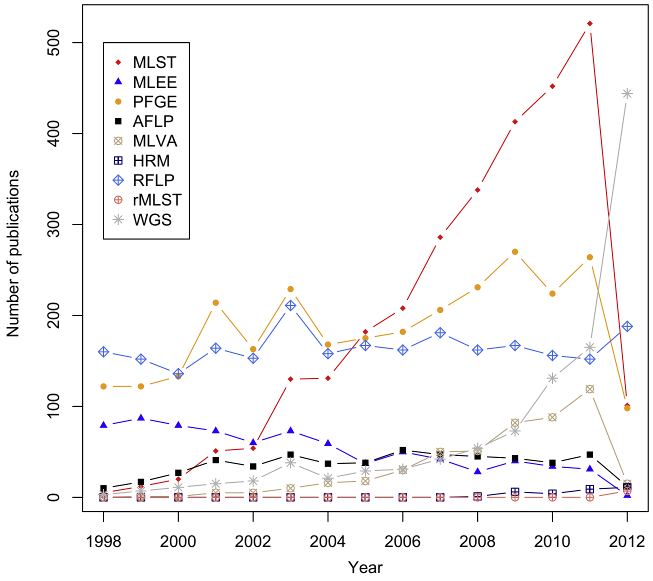

Fig. 1. Number of publications related to bacterial typing methods as a function of time. Abbreviations are defined in Section 1. WGS = whole-genome sequencing.

Sequence Database (BIGS$_{DB}$), which implements a combined taxonomic and typing approach for the whole domain of bacteria and the analysis of linked phenotypic and genotypic information (Jolley and Maiden, 2010). More recently, a Bayesian model-based method also offers the possibility to automatically relate unidentified isolates with information deposited in curated databases (Cheng et al., 2011). This method can be used with any MLST dataset through the software BAPS 5.4 (http://www.helsinki.fi/bsg/software/).

Given the success of website technologies, recent efforts have exploited the potential of Internet resources to incorporate geospatial information in bacterial epidemiological studies (Aanensen et al., 2009; Baker et al., 2010; Grundmann et al., 2010). The websites www.spatialepidemiology.net/ and beta.mlst.net/Instructions/mlstmaps.html, for example, provide precise locality data related to strain distribution and also provide 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.

1.2. The MLST scheme: a comparison with other bacterial typing methods

To be useful, a strain typing method should provide enough discriminatory power to distinguish between isolates from unlinked sources and to be sufficiently reliable to cluster isolates from the same source (Killgore et al., 2008; Unemo and Dillon, 2011). Since its proposal in 1998, MLST rapidly emerged as the state-of-the-art technique for bacterial molecular typing over other techniques (Fig. 1). Unfortunately the MLST scheme is not the panacea to address all questions pertaining to molecular epidemiology, and alternative methods exist that offer complementary or even better discriminatory power at different temporal scales (see Table 1). In addition to this, the cost issue is also pivotal when choosing a bacterial typing technique and a considerable number of isolates need to be investigated.

Currently, the main drawback of the MLST method is that the selection of housekeeping loci requires a reference genome (Parkhill et al., 2003; Sreevatsan et al., 1997). Moreover, the lack of diversity throughout entire genomes or housekeeping genes in some pathogens, as well as the presence of recently emerged species or recent population bottlenecks, may yield the MLST scheme very limited in discriminatory power (Harbottle et al., 2006; Pourcel et al., 2004; Torpdahl et al., 2005). Until the development of MLST, the most widely used technique for indexing allelic variation was the multilocus enzyme electrophoresis approach (MLEE). A major drawback of the MLEE is that only genetic changes altering the electrophoretic properties of the studied protein can be detected (about one 20th of all possible mutations), and consequently synonymous mutations are overlooked. Alternative gel-based methods, such as the pulsed-field gel electrophoresis (PFGE), restriction fragment length polymorphism (RFLP) or amplified fragment length polymorphism (AFLP) offer a more affordable alternative to the MLST scheme and can provide better resolution at short-temporal scales in some bacterial species (Melles et al., 2007). However, the MLST approach is usually preferred because in all these gel-based approaches, comparison of results between laboratories is often problematic and a high level of expertise is needed to interpret and to translate banding patterns.

Another multiple-locus technique is the variable number of tandem repeats analysis (MLVA), which is based on the analyses of polymorphic repeated sequences (VNTR). Comparative studies between MLVA and MLST have yielded similar results (Elberse et al., 2011; Malachowa et al., 2005; Schouls et al., 2006; Top et al., 2004), and in recently originated species, the MLVA approach has higher discriminatory power (Vergnaud and Pourcel, 2006). This technique shares all the advantages of the MLST scheme in terms of portability and reproducibility at a lower cost, but VNTR may evolve too quickly to provide reliable phylogenetic relationships among closely related strains and the size difference may not always reflect the real number of tandem repetitions because the presence of insertions and deletions (Li et al., 2009).

Recently a new methodology has been proposed based on high resolution melting curves (HRM) to distinguish single base variation and so identify SNPs without the burden of sequencing (Erali et al., 2008; Taylor, 2009). After amplification, PCR products are characterized in relation to their disassociation (melting) curves. This method provides a rapid, close-tubed, highly efficient and low cost strategy for detecting base substitutions and small insertions or deletions (Millat et al., 2009). However, the detection of an unidentified melting profile demands sequencing to identify the new profile and thus an increasing cost.

In addition, the ribosomal multilocus sequence typing method (rMLST) has been proposed to index the molecular variation of 53 genes encoding bacterial ribosome protein subunits (Jolley et al., 2012a). This novel method pursues the integration of a taxonomic and typing method in a similar curated MLST scheme. Data generated can be easily accessed and accommodated in the abovementioned database BIGS$_{DB,}$ a reference genome is not required, targeted loci are conserved across the whole bacteria domain and the reanalysis of existing allele designations is not required (Jolley and Maiden, 2010). Although more expensive, the rMLST is likely to provide better resolution than previous methodologies, which coupled with the decreasing cost of sequencing DNA make it a promising technique. The method still requires further exploration, but certainly it has the potential to provide a universal bacterial typing method extending the idea of the MLST scheme.

Finally, in order to achieve greater resolution, a method has been developed that relies on presence or absence of pan-genomic or distributed genes among bacterial species that have the same MLST profile. This clustering method leverages the massive amount of whole genome information that is being accumulated and has utility in resolving close strain relationships (Hall et al., 2010). This novel method represents an affordable technique if

Table 1
Comparison of most common bacterial typing techniques (adapted from Foxman et al., 2005). See Section 1 for abbreviations referred to typing methods.

Typing method	Method description	No. of markers	Temporal scale	Variation source	Discriminatory power	Reproducibility	Equipment/time	Equipment/consumables-reaction costs (per isolate)	Available databases
MLST	PCR amplification of housekeeping genes to create an allelic profile	7	Macroepidemiological Microepidemiological	DNA sequence	Moderate to high	High	Thermal cycler/ moderate	$30–45 K High ~$80	pubmlst.org www.mlst.net mlst.ucc.ie www.pasteur.fr/ mlst
MLEE	Phenotypic characterization of the electrophoretic mobility of housekeeping enzymes	10–20	Macroepidemiological Microepidemiological	Electrophoretic mobility	Moderate	Moderate	Gel box, switching unit, cooler, power supply/moderate	$10–20 K Moderate ~$20	NA
PFGE	Comparison of large genomic DNA fragments after digestion with rare restriction enzyme	NA	Microepidemiological	Banding pattern	Moderate to high	High	Gel box, switching unit, cooler, power supply/high	$10–20 K Moderate ~$22	NA
AFLP	Digestion of genomic DNA with two restriction enzymes, ligation of restriction fragments and selective amplification	NA	Microepidemiological	Banding pattern	Moderate to high	Low	Thermal cycler/ moderate	$8–12 K Moderate ~$20	NA
MLVA	PCR amplification of VNTR loci followed by sizing of the PCR products to create an allelic profile	10–80	Microepidemiological	DNA sequence	Moderate to high	High	Thermal cycler/low	$30–45 K Moderate ~$20	minisatellites.u-psud.fr www.mlva.net www.pasteur.fr/ mlst
HRM	PCR amplification followed by characterization of amplicon melting curves	NA	Macroepidemiological Microepidemiological	Melting temperature	High	High	Real time thermal cycler/very low[a]	$30–45 K Very low[a]	NA
RFLP	Digestion of genomic DNA with restriction enzimes to produce multiple short restriction fragments	NA	Microepidemiological	Banding pattern	Low	Low	Southern transfer/ high	$8–12 K Low ~$14	NA
rMLST	PCR amplification of *rps* genes to create an allelic profile	53	Macroepidemiological Microepidemiological	DNA sequence	High	High	Thermal cycler/ moderate	$30–45 K High ~$600 (if WGS is not needed)	http://pubmlst.org/ software/database/ bigsdb/
Pan-genome	Detection of similarities/differences in the pan-genomic or distributed genes	>1000	Macroepidemiological Microepidemiological	Presence/ absence of genes	High	High	NGS platforms or microarrays/ moderate to high	$80–130 K Very high ~$1–20 K per run depending on the NGS platform used	www.francisella.org

NA = not applicable.
[a] If new melting profiles are not detected.

microarrays are used; however, the cost of this approach increases dramatically if whole genome sequencing is required.

1.3. Population genetics and phylogenetics under the MLST scheme

The MLST scheme was originally proposed for the identification of highly related bacterial genotypes (Maiden et al., 1998), but the genealogical information inherited in the DNA sequences also allowed one to address questions about species boundaries, population dynamics, and evolutionary relationships (Spratt, 1999). Different mechanisms for the exchange of genetic material among bacteria were known for years (Lorenz and Wackernagel, 1994), but their role on population structure was widely assumed to be negligible. This paradigm radically changed after several studies revealed extensive genetic exchange caused by recombination (e.g., DuBose et al., 1988), which entailed a broad spectrum of bacterial populations ranging from fully clonal (recombination does not effectively occur) to non-clonal populations (genetic diversity is randomized by frequent events of genetic exchange). Subsequent evidence showed that those extremes are rare in nature, and most bacterial population exhibit high levels of recombination, but not sufficient to prevent the emergence of clonal lineages (Spratt, 1999).

With population genetic and phylogenetic studies of bacterial species, then, one is forced to examine the role of genetic recombination (Posada et al., 2002). In this regard, the MLST scheme tries to overcome this problem by combining several neutral molecular markers scattered across the genome that are relatively short in length, thereby avoiding complications due to recombination (Maiden et al., 1998). As in all population studies, the sampling strategy is critical to avoid bias towards certain isolates and to accurately assess the overall genetic variation in the population. Housekeeping genes usually offer enough resolution to accurately infer population parameters and reconstruct phylogenetic relationships. However there is no single core of universal genes that can be used throughout all pathogens (but see Jolley et al., 2012a), since recombination, substitution and selection rates vary across loci and species (Pérez-Losada et al., 2006); therefore, choosing the appropriate set of loci ultimately relies upon the biology of the individual species under study (Spratt, 1999). Molecular phylogenetic studies based on microbial populations face problems that are not often encountered in typical evolutionary studies (Fraser et al., 2007). Bacterial species typically exist as clusters of genetically related strains (Acinas et al., 2004), but finding those clusters may not be straightforward since high rates of recombination can certainly render meaningless and misleading phylogenetic trees (Posada and Crandall, 2002). In addition, isolates tend to be very closely related and frequently both the parent strains and their descendants are included in the same sample (Hall and Barlow, 2006). Thus, recombination requires a different paradigm for visualizing genealogical relationships as networks instead of trees (Posada and Crandall, 2001) and special approaches for estimating population genetic parameters that accommodate the biological reality of recombination (Schierup and Hein, 2000).

1.4. Housekeeping genes: diversity levels and phylogenetic resolution

The MLST approach uses only a small fraction of the genome (usually between 6 and 7 housekeeping genes of approximately 450–500 bp), which is assumed to be a representative sample of the entire genome diversity (Didelot and Maiden, 2010). Protein-encoding housekeeping genes are viewed as the most reliable markers, since they are presumed to evolve slowly by the random accumulation of neutral variation, providing much more reliable data for both accurate typing and phylogeny estimation.

Levels of genetic polymorphism in housekeeping genes are usually high enough to assess population structure and strain relatedness (Maiden, 2006). However, how much genetic variability is necessary to accurately infer inter- and intra-species evolutionary relationships remains an open question; similarly, the correlation between gene function and phylogenetic resolution has been barely addressed (Cooper and Feil, 2006; Ferreira et al., 2012; Zeigler, 2003). For example, contrarily to expectations, Kuhn et al. (2006), Robinson et al. (2005) and Cooper and Feil (2006) showed for *Staphylococcus aureus* that the inclusion of rapidly evolving genes under diversifying selection did not hamper the accurate inferences of evolutionary parameters (Cooper and Feil, 2006; Kuhn et al., 2006; Robinson et al., 2005); in fact, in the same studies, standard MLST genes provided the poorest phylogenetic resolution. These results suggested that loci selection, at least at the intra-species level, should be primarily based on nucleotide diversity rather than gene function (Cooper and Feil, 2006). Hence, if higher resolution is required, including more fast-evolving genes (as those subject to positive diversifying selection) might be more beneficial than adding more MLST genes (Maiden, 2006).

It is not clear what values of genetic variability yield better phylogenetic estimates or why variation greater than 1% generally does not improve resolution (Cooper and Feil, 2006). As a general rule, it has been suggested that loci comprising at least the average diversity for all genes may have the potential to accurately trace molecular epidemiological studies (Cooper and Feil, 2006). The presence of "sufficient diversity" is a critical factor when analyzing closely related strains within species. This issue becomes less problematic at higher taxonomic levels, and in that case, MLST data are likely to provide the appropriate framework for studying molecular epidemiology in microbial pathogens. Several studies have tried to identify a universal set of housekeeping genes for bacterial typing and prediction of phylogenetic relatedness at different taxonomic levels (Stackebrandt et al., 2002; Zeigler, 2003). These studies have shown that a careful selection of single genes could be sufficient for discriminating between bacterial species, but the inference of intrageneric evolutionary relationships may be difficult when a small set of genes is used (Zeigler, 2003).

More recently, cutting-edge approaches based on full-genome sequences have been applied with the expectation that including more genetic data will buffer the effect of non-informative loci (Schürch and van Soolingen, 2012). However, Ferreira et al. (2012) have pointed out the need for a careful examination of genomic features such as polymorphism dispersion, intergeneric region sizes, and positively selected loci ratios; since these factors may impact recombination and mutation rates differently, resulting in non-convergent and incongruent phylogenies. In agreement with previous studies, Ferreira et al. (2012) also showed that inclusion of positively selected genes did not prevent the accurate inference of the evolutionary parameters, and curiously, non-coding regions yielded similar results. Although this study relates to a specific bacterial species, it provides valuable clues about the potential of non-standard loci as potential markers for MLST. Currently, most inferences on bacterial evolution have been and are still produced using MLST data. However, next-generation sequencing platforms now provide the means to capture multiple non-standard target loci to detect single nucleotide polymorphisms or to sequence full genomes. Such methods are briefly described in the next section.

2. Sequencing approaches to MLST

Next-generation sequencing (NGS) is permeating many aspects of biology including those endeavors typically related to MLST (Metzker, 2010). Although traditionally Sanger sequencing is still used more than NGS, as revealed by a simple Web of Science search

(Sanger/NGS = 2 × 10^6/5 × 10^4), the latter is gaining popularity for reasons such as affordability (when sequencing large numbers of samples), scalability, and marker discovery (gene mining). In this section, we present a review of the sequencing approaches currently used in relation to MLST.

2.1. Sanger sequencing

Traditional Sanger sequencing still enjoys great popularity primarily because of its low cost at small scales and perceived superior quality when it comes to error rates (Hoff, 2009). In a nutshell, to carry a Sanger reaction we need a single stranded DNA molecule plus dideoxy-nucleotides triphosphates (along with tagged chain terminators) and a primer that will be extended by a DNA polymerase. Tagged amplicons of different lengths are then fractionated via electrophoresis or with a chromatography capillary column so that "color" tags are read and a digital consensus sequence is inferred. Sanger sequencing provides unambiguous DNA sequence markers that can be used to design MLST schemes. Read lengths, or the mean/mode length achieved by a sequencing method, are typically longer in Sanger than those generated by other sequencing approaches, which may reduce the number of loci required for accurate bacterial characterization. Additionally, Sanger is amenable to sequencing single molecules and therefore reduces the potential impact of artificial recombination; implying that all detectable recombinant signals come from real biological events (Salazar-Gonzalez et al., 2008). Moreover, post-processing in Sanger sequencing is simple compared to NGS, which lends itself to be preferentially used in laboratories lacking strong bioinformatic capabilities.

Sanger sequencing is still the gold standard for generating DNA sequence data (Harismendy et al., 2009). One of its more attractive features is its low error rate (from 0.0001% to 1%), which seems to depend on the algorithms used for post-processing (Ewing and Green, 1998; Ewing et al., 1998). NGS techniques such as pyrosequencing, on the other hand, report error rates of 0.49–2.8% (Harismendy et al., 2009), though the technologies are improving regarding sequencing chemistry and software.

2.2. Next-generation sequencing

Although Sanger sequencing still can fulfill the needs of many microbiology labs, the prospects that NGS technologies offer, along with the dimensions of their benefits, will likely surpass Sanger sequencing (Castro-Nallar et al., 2012). Large-scale sequencing efforts using Sanger require expensive infrastructure and laborious bench work (Medini et al., 2008). Several platforms and chemistries are available within NGS; however, large-scale projects can be done on a bench-top machine with ease (see Hui, 2012 for a review).

NGS contributes at least twofold to the development of MLST schemes. First, traditional MLST schemes need a reference genome in order to develop appropriate markers (Table 2). Currently, there are many genomes available from which one can extract marker information (3.334 complete and 11.056 incomplete; GOLD database; http://www.genomesonline.org). In fact, software implementations such as PhyloMark (http://sourceforge.net/projects/phylomark/) are already accessible and can aid with genome-wide marker examinations. The aim of PhyloMark is to identify the minimum number of markers that recapitulate a full genome phylogeny (Sahl et al., 2012) (Fig. 2B). Due to NGS technologies, the number of available bacterial genomes is increasing at a fast pace. However, still a large proportion of bacteria are lacking genome information and thus the abovementioned strategies cannot be applied. Secondly, NGS has proved useful in generating sequence data when little is known about the target organisms by providing the raw material to extract markers for MLST schemes (Fig. 2C). Furthermore, NGS read lengths are now falling within the size range of the genes (450–500 bp) used in MLST (http://454.com), and with the addition of multiplexing IDs (MID), it is possible to pool large numbers of samples and still get the benefit of sequencing sample targets with high depth (coverage).

Sanger sequencing is a mature technology with little room to improve. In contrast, NGS technologies are rapidly evolving in a complementary non-overlapping manner. For instance, pyrosequencing is improving both regarding homopolymer errors and read lengths. On the other hand, Illumina systems do not provide reads as long as those from pyrosequencing, but its coverage is greater, which could be advantageous for assessing bacterial genetic diversity in intra-host dynamics. A combination of high density short-read technologies (e.g., Illumina) with long read (but relatively low accuracy) third-generation direct reads (e.g., Pacific Biosciences) and novel assembly algorithms suggests a productive approach for future bacterial genome sequencing and assembly (Ribeiro et al., 2012).

To date, several methodologies have been put forward to improve traditional MLST schemes, many of which are taking advantage of NGS. In general, they fall into a category in which, given presence or absence of genetic information, they use some sort of gene/genomic region targeting (Fig. 2A and B) or enrichment

Table 2
Comparison of NGS-based methods used in gene mining and sequencing.

Method	Sequencing technology	Read length (bp)	Genetic markers needed	Reference genome needed	Library preparation	Good for mining	No. of markers[c]	Bioinformatic post-processing/software
TAS	454	400–800[a]	Yes	Yes	No	No	6 Genes for >44 taxa	Medium/ barcodecruncher
HiMLST	454	400–800	Yes	Yes	No	No	7 Genes for >575 taxa	Medium/roche
Anchored hybrid/ultra conserved elements enrichment	Illumina	100	No	Yes[b]	Yes	Yes	512/854	High/open or free software[d]
PRGmatic	454	400–800	No	No	No	Yes	780 Genes for >20 taxa	High/open or free software[d]
Traditional MLST	Sanger	800	Yes	Yes	No	No	7 Genes	Little/open or free software[d]

[a] Read lengths reported on www.454.com using the new GS FLX + system.
[b] Although no reference is strictly needed, closely related genomic sequences are necessary in order to design appropriate probes.
[c] Numbers based on figures reported on original papers.
[d] For details about software used check original papers.

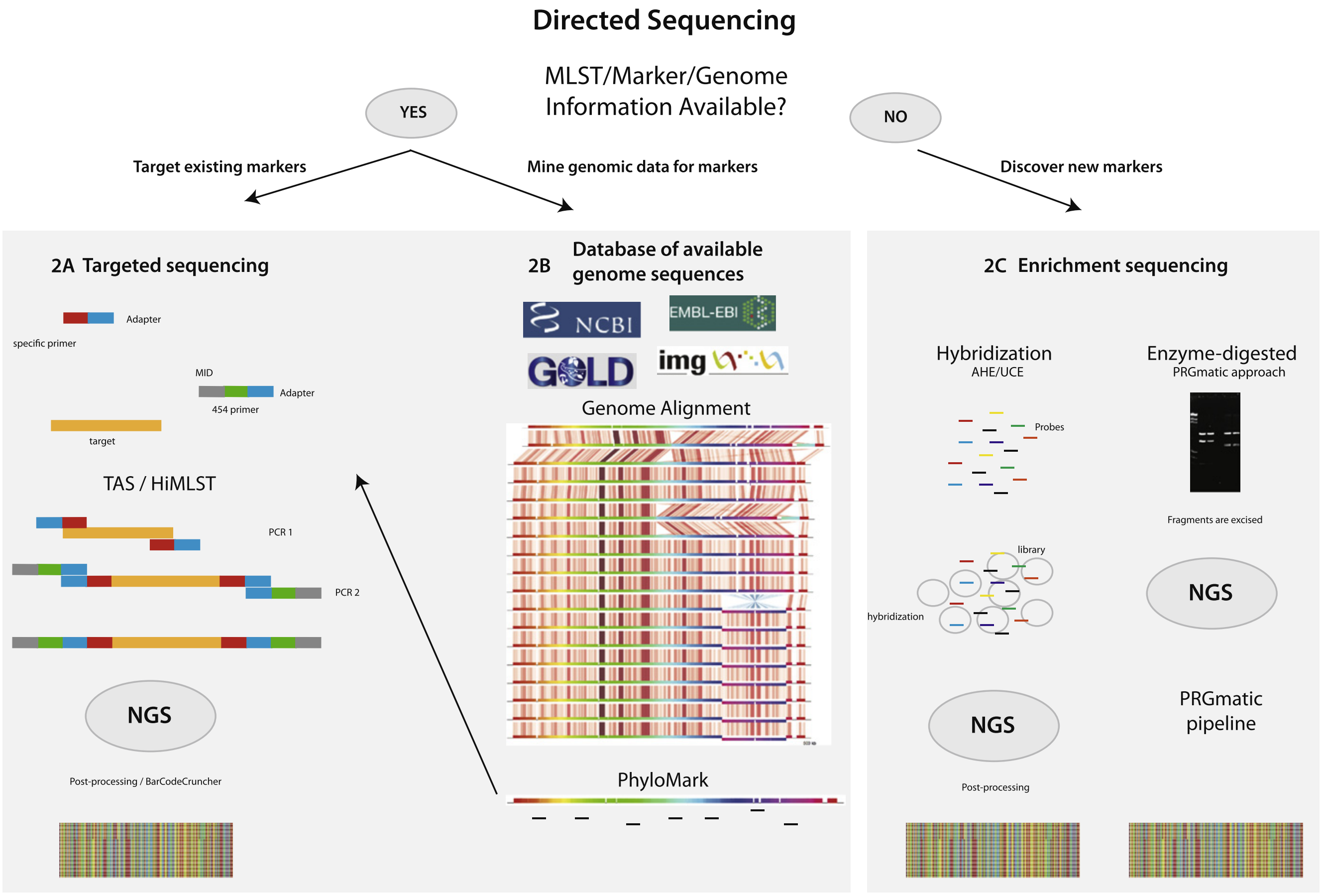

Fig. 2. Schematic diagram showing direct sequencing approaches to obtain and discover genetic markers for MLST analysis. Left (A and B) and right (C) panels show approaches when genomic information is available or not, respectively. TAS = targeted-amplicon sequencing; HiMLST = high-throughput MLST; AHE = anchored hybrid enrichment; UCE = ultra-conserved elements enrichment. See Section 2 for other abbreviations and further detail.

(Fig. 2C) to obtain potential marker sequences that can be used in downstream MLST applications. If genomic information of the group of interest is available, it is possible to develop markers that would resemble genomic relationships (Fig. 2B). In turn, if no information is available except from related taxa, then it is possible to design sequence capture experiments (usually with probes) to develop or discover new markers (Fig. 2C). Alternatively, if no genomic information exists for the group of interest, an enriched *de novo* approach can be also applied to discover new markers.

With the decreasing cost of NGS, new affordable applications have arisen to perfect or create new ways of generating and analyzing sequence-typing data. A natural step toward high-throughput sequence typing is to combine the power of NGS with sequence targeting for which some extent of variability is already known. Methods relying on targeting known genes (Fig. 2A and B) or enriching genomic fractions to discover new markers (Fig. 2C) are now available (Table 2). In general, these methods, though not heavily used yet, promise to overcome some of the limitations of the MLST classic approach. For instance, the lack of a reference genome might not be a limitation since by performing enrichment steps prior to NGS, it is possible to single out large homologous regions of the genome that can be scaled up to analyze larger datasets and/or more populations.

One example is the targeted-amplicon sequencing method (TAS; Fig. 2A), which capitalizes on NGS to sequence a large number of regions from large numbers of pooled samples (Bybee et al., 2011). Given its relatively longer reads (800 bps) compared to other NGS technologies, pyrosequencing has been the preferred choice of targeted approaches. Recently, a method was made available in which MLST genes are amplified in a two-step PCR using sequence specific primers that have attached MID (HiMLST), similar to what it is routinely done when adding a restriction site to a target gene (Boers et al., 2012). Then, samples from multiple strains or species are pooled and sequenced as usual in a 454 Roche machine. It is worth noting that Roche 454 technology is able to deliver reads of up to 800 bp (using the GS FLX + system), which may be particularly useful for MLST analysis (www.454.com). This method is essentially the same as the TAS method published earlier but specifically designed for MLST. Both approaches use MID multiplexing capabilities, so costs are lowered by pooling samples. A simple post-processing procedure guarantees that sequences are obtained in a per strain/species basis, for example by using the BarcodeCruncher software (http://crandalllab.byu.edu/Computer-Software.aspx). The reported HiMLST protocol was able to profile 575 isolates from several bacterial species (7 genes). In addition, the TAS protocol was able to obtain sequences from 6 genes over 44 taxa in a quarter plate (Table 2; Boers et al., 2012; Bybee et al., 2011).

On the other hand, examples of directed sequencing by enrichment are: (1) Anchored Hybrid Enrichment/Ultra conserved Elements (Faircloth et al., 2012; Lemmon et al., 2012) and (2) the PRGmatic approach (Hird et al., 2011). Although these methods have been originally developed for phylogenomics and high-level systematics (i.e., phylogenies of species), they can also be applicable to MLST, since multiple informative markers are also often needed to resolve genealogical relationships among individuals. Enrichment methods (or sequence capture methods) can be of help when little is known about the species under scrutiny or the objective is to discover new MLST markers. The PRGmatic approach, for example, uses restriction enzyme-digested, size-selected genomic DNA sequenced by pyrosequencing. Then, it clusters aligned reads by identity into alleles and then into loci. A great innovation of the method is that it generates a provisional reference genome (PRG) that is further used to align reads and generate sequences for each locus. In turn, the anchor hybrid enrichment method (or ultra conserved elements enrichment by Faircloth et al., 2012), probably a more powerful approach in terms of finding loci, attempts to "capture" conserved genomic regions using probes and then sequence them using the Illumina platform. The post processing is fairly straightforward in terms of bioinformatic burden, though trained personnel are probably necessary to automate post-processing by writing tailored computer scripts. Although this method is more powerful regarding the number of loci recovered, it is likely to be more expensive as well. In particular, DNA library generation could be an economic burden for a medium-sized laboratory in terms of initial investment (Table 2). However, per base or per loci sequencing costs are very low compared to other NGS-based methods. Other enrichment methods are discussed elsewhere (Cronn et al., 2012; Mamanova et al., 2010).

In principle, due to their higher sequencing power (up to 854 loci; Table 2), all the above mentioned approaches should help to overcome some of the problems standard MLST schemes may encounter, such as lack of diversity in genome or housekeeping genes, or more importantly, the ability to detect patterns in emergent species or in species under demographic or selective processes. Very few studies looking at bacterial evolution and epidemiology using these methodologies have been published so far. As sequencing costs keep decreasing, we foresee an increase in MLST studies using NGS. Coupling NGS to MLST is a challenge and new strategies are starting to emerge. Recently, for example, Singh et al. (2012) developed a hairpin-primed multiple amplification method that can amplify numerous target genes simultaneously.

3. Analysis of MLST

Methods of analysis of MLST data can be classified in two basic strategies: (a) those that rely on allele and ST designations to estimate relatedness among isolates (*allele-based methods*) and so ignore the number of nucleotide differences between alleles; and (b) those that rely on nucleotide sequences directly to estimate relatedness and population parameters (*nucleotide-based methods*) (Table 3). The allele-based approach is thought to work well in non-clonal organisms (e.g., *Helicobacter pylori*), while nucleotide-based approaches are preferable for clonal organisms (e.g., *S. aureus*), since the former approaches are likely misleading because they cannot distinguish between single-base changes in multiple loci versus multiple mutations in the same number of loci (Maiden, 2006). In practice, most microbes show some degree of clonality (clonal complex) in their populations, hence, in principle, both types of analyses could be carried out in population and epidemiological studies (e.g., Tazi et al., 2010).

3.1. Allele-based methods

These types of methods require first the coding of DNA sequences from each locus into numbers using information available in public MLST databases (see Section 1). If no match is found, a new number is assigned in order of discovery. Several computer programs, such as sequence typing analysis and retrieval system (STARS), have been developed for this task. Once alleles have been assigned, data are entered in the MLST databases to acquire an ST

Table 3
List of population genetics programs listed in this review including their functionalities and online links.

Data type	Functionality	Link
Alleles		
STARS	Allele assignment	http://sara.molbiol.ox.ac.uk/userweb/mchan/stars/
START2	Data summary/exploratory analysis	http://pubmlst.org/software/analysis/start2/
eBURST	Inference of patterns of evolutionary descent	http://eburst.mlst.net/
goeBURST	Inference of patterns of evolutionary descent using matroids	http://goeburst.phyloviz.net/
PHYLOViZ	Inference of patterns of evolutionary descent	http://www.phyloviz.net/wiki/
Nucleotides		
Phylogenetics		
MAFFT	Sequence alignment	http://mafft.cbrc.jp/alignment/software/
MAUVE	Sequence alignment	http://asap.ahabs.wisc.edu/mauve/index.php
JModeltest2	Selection of models of nucleotide substitution	https://code.google.com/p/jmodeltest2/
RAxML	ML inference of evolutionary relationships	http://www.exelixis-lab.org/
GARLI	ML inference of evolutionary relationships	http://code.google.com/p/garli/
PHYML	ML inference of evolutionary relationships	http://code.google.com/p/phyml/
MrBayes	Bayesian inference of evolutionary relationships	http://mrbayes.sourceforge.net/
BEAST	Bayesian inference of evolutionary relationship	http://beast.bio.ed.ac.uk/Main_Page
ClonalFrame	Bayesian inference of clonal relationships considering recombination	http://www.xavierdidelot.xtreemhost.com/clonalframe.htm
UMP	Inference of reticulated evolutionary relationships	http://applications.lanevol.org/combineTrees/
TCS	Inference of reticulated evolutionary relationships	http://darwin.uvigo.es/software/tcs.html
SplitTrees4	Inference of reticulated evolutionary relationships	http://www.splitstree.org/
BEST	Coalescent inference of gene and species trees	http://www.stat.osu.edu/~dkp/BEST/introduction/
*BEAST	Coalescent inference of gene and species trees	http://beast.bio.ed.ac.uk/Main_Page
Population dynamics		
BEAST	Coalescent inference of population parameters, demography and divergence times	http://beast.bio.ed.ac.uk/Main_Page
LAMARC	Coalescent inference of population parameters, demography and divergence times	http://evolution.genetics.washington.edu/lamarc/index.html
PAML	ML inference of population parameters and phylogenies	http://abacus.gene.ucl.ac.uk/software/paml.html
OmegaMap	Bayesian inference of population parameters	http://www.danielwilson.me.uk/omegaMap.html
HYPHY	ML and Bayesian inference of population parameters and phylogenies	http://hyphy.org/w/index.php/Main_Page
SPREAD	Bayesian phylogeography	http://www.kuleuven.ac.be/aidslab/phylogeography/SPREAD.html

profile. At this point exploratory analysis (e.g., allele and profile frequencies, polymorphism estimates, codon usage, etc.) could be performed using sequence type analysis and recombinational tests (START2) software (Jolley et al., 2001). Relatedness among STs can then be displayed using methods of cluster reconstruction such as the simple unweighted pair group method with arithmetic mean (UPGMA) and the based upon related sequences types (eBURST) approach. The former method uses a matrix of distances among STs to estimate isolate relatedness, while eBURST (Feil et al., 2004) infers patterns of evolutionary descent among isolates using a simple model of clonal expansion and diversification. A new globally optimized version (goeBURST) has also been developed that identifies alternative patterns of descent using graphic matroids (Francisco et al., 2009). Recently, a new approach (PHYLOViZ) has been 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 (Francisco et al., 2012). PHYLOViZ uses goeBURST for representing the possible evolutionary relationships between strains.

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. They are, therefore, better suited for exploratory data analysis rather than fine statistical inference (Didelot and Falush, 2007). A larger and more sophisticated plethora of nucleotide-based methods exist to estimate isolate relationships and population parameters.

3.2. Nucleotide-based methods

Any analysis of nucleotide data usually begins with an alignment (i.e., estimation of site homology; Rosenberg, 2009). Several fast and accurate strategies for aligning gene regions and genomes are implemented in MAFFT (Katoh et al., 2005) and MAUVE (Darling et al., 2010), respectively. After the alignment has been generated, we need to determine the model of evolution that fits the data the best. Model choice is a critical issue and the chosen model (or lack thereof) will affect all subsequent phylogenetic (Section 3.2.1) and population (Section 3.2.2) analyses (Kelsey et al., 1999). This issue is usually assessed within a maximum likelihood or Bayesian phylogenetic framework and under multiple criteria, like the Akaike or Bayesian Information Criterion and marginal likelihoods (see Baele et al., 2012; Posada and Buckley, 2004; Xie et al., 2011). These and other model choice strategies are implemented in JModeltest2 (Darriba et al., 2012).

3.2.1. Phylogenetic relatedness

Phylogenetic reconstruction methods can be divided into two types, those that proceed algorithmically (e.g., UPGMA, Neighbor-joining) and those based on optimality criteria. Here we will focus on the latter since we find this feature particularly important for analyzing MLST data; a more extensive review of phylogenetic methods can be found in Pérez-Losada et al. (2007a).

Maximum likelihood (ML) inference attempts to identify the topology that explains the evolution of a set of aligned sequences under a given model of evolution with the greatest likelihood (Felsenstein, 1981). RAxML (Stamatakis, 2006), GARLI (Zwickl, 2006) or PHYML (Guindon et al., 2010) implement the ML criterion efficiently and accurately and can handle datasets of >1.000 sequences. Confidence in the estimated ML relationships (i.e., clade support) can be assessed using the nonparametric bootstrap procedure (Felsenstein, 1985).

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 that the tree(s) is (are) correct (Huelsenbeck et al., 2001). BI has advantage over ML approaches both in accounting for uncertainty in the phylogeny and model parameters estimated, and allowing for hypothesis testing. Clade support is estimated by summarizing the frequency of that clade across a distribution of trees through a consensus analysis. Bayesian phylogenies are estimated using Metropolis-coupled Markov chain Monte Carlo (MC^3) methods and both are implemented in programs like MrBayes (Ronquist and Huelsenbeck, 2003) or BEAST (Drummond and Rambaut, 2007). The output generated by these programs can then be evaluated in Tracer (Rambaut and Drummond, 2009) to confirm that MC^3 chains have mixed well and converged.

Standard phylogenetic methods assume a lack of recombination, an assumption violated by many microorganisms. Hence if recombination is suspected in our data, we should first detect and eliminate recombinant regions or identify breakpoints (see Section 3.2.2 below), so alignments can then be subdivided into non-recombinant regions and analyzed separately. Alternatively, one could use an approach that takes homologous recombination into account while inferring clonal relationships between the members of a sample. Such a method is implemented in ClonalFrame (Didelot and Falush, 2007) within a Bayesian coalescent framework. Similarly, phylogenetic strategies that assume a reticulated model of evolution (network) instead of a bifurcating tree may be better when recombination is substantial (Posada and Crandall, 2001); the Union of Maximum Parsimonious trees (Cassens et al., 2005) and TCS (Templeton et al., 1992) are two of such approaches and both perform well under relatively low levels of diversity and recombination (Woolley et al., 2008). Another broadly used network approach is SplitsTree4 (Huson and Bryant, 2006). An interesting application of the network strategy has been recently developed by Plucinski et al. (2011) to infer local and global properties of the host populations in commensal bacteria.

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., incomplete lineage sorting). New coalescent methods have been developed to deal with stochastic variation in gene trees. Among these, the Bayesian-based BEST (Liu, 2008), STEM (Kubatko et al., 2009), and $*$BEAST (Heled and Drummond, 2010) approaches are well suited to estimate the joint posterior distribution of gene trees and the organism tree using multilocus molecular data.

3.2.2. Population dynamics

The evolution of DNA sequences in natural populations can be described by parameters like 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, 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. Comprehensive reviews of statistical methods for detecting and estimating recombination rates are presented in Posada et al. (2002); although since then, new methods have been developed (e.g., Jeffrey, 2004; Lefebvre and Labuda, 2008; Padhukasahasram et al., 2006; Wang and Rannala, 2008, 2009) and revised (e.g., Auton and McVean, 2012; Martin et al., 2011; Stumpf and McVean, 2003). Posada et al. (2002) concluded that multiple methods should be used to detect or estimate recombination. Consequently, software packages like RDP4 (Martin et al., 2010) have been developed to implement up to eight recombination estimators that allow the user to draw conclusions based on the outcome of multiple tests.

Genetic diversity is the most important population parameter and is usually estimated in relation to recombination as the rate of recombination to mutation (r/m), so the relative impact of each force on generating microbe genetic diversity can be assessed (Feil et al., 1999). Reviews of classical and coalescent statistical methods for estimating genetic diversity can be found in Pearse and Crandall (2004), Excoffier and Heckel (2006) and Waples and Gaggiotti (2006); nonetheless newer methods have been developed since these reviews (e.g., Bashalkhanov et al., 2009).

Growth rates reflect the variation of genetic diversity along time. Growth can be estimated under a certain demographic model (e.g., exponential) or without dependence on a pre-specified model, such as the Bayesian skyline plot (Drummond et al., 2005) or the Skyride model (Minin et al., 2008), both implemented in BEAST. Interestingly, BEAST also allows for the analysis of temporally spaced sequence data. Recombination, genetic diversity, and exponential growth rates can all be estimated in LAMARC (Kuhner, 2006).

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 $\omega \approx 0$ lack of selection. ω is usually estimated within a ML phylogenetic framework and assuming an explicit model of codon substitution. If significant evidence (usually obtained through likelihood ratio tests) of adaptive selection is obtained, then Bayesian tests can be applied to detect amino acid sites under selection (e.g., Yang et al., 2005). These methods are implemented and described in more detail in PAML (Yang, 2007). If recombination is present, other methods exist that can estimate recombination and selection rates simultaneously (OmegaMap; Wilson and McVean, 2006), or account for the former while estimating the latter (HYPHY; Kosakovsky Pond et al., 2005).

Other key factors in pathogen dynamics are the time of emergence of the epidemic and the geographical distribution of pathogens. New probabilistic models have been recently developed within the Bayesian framework (Lemey et al., 2009, 2010) that allow the inference and hypothesis testing of divergence times, ancestral locations and historical patterns of migration (i.e., phylogeographic history). Those parameters can be estimated in BEAST and SPREAD (Bielejec et al., 2011) and visualized using virtual globe software like Google Earth (www.google.com/earth/index.html). Such methods have already begun to be applied to the analysis of MLST and/or genome and SNP (see Section 5) data (Gray et al., 2011; McAdam et al., 2012; Weinert et al., 2012). Similarly, divergence times and ancestral states can be also estimated in LAMARC.

4. Applications of MLST

The popularity of MLST is driven by its ease of use and discriminating power. Consequently, over the last few years we have seen not only an increase in MLST schemes (Fig. 1) and sequence types available, but also in the diversity of their applications. Although primarily developed for pathogen identification (typing), MLST sequence data have also been applied to other aspects of molecular epidemiology (e.g., disease transmission, evolution of virulence) and public health (e.g., monitor vaccination programs), as well as to other areas such as phylogenetics, taxonomy, speciation, population genetics, biosafety, and even to the inference of human migrations. Below we list a series of examples taken from the most recently published literature showing some of those applications.

4.1. Molecular epidemiology and public health

MLST has become the routine typing approach for the identification of clinical specimens. Accurate and quick characterization of organisms is crucial for epidemiological surveillance (Brehony et al., 2007; Trotter et al., 2007), detection and management of disease outbreaks (Byrnes et al., 2010; Palazzo et al., 2011; Vanderkooi et al., 2011), estimate prevalence rates (Haran et al., 2012; Ibarz-Pavon et al., 2011; Sproston et al., 2011) or study horizontal (Stensvold et al., 2012; Walker et al., 2012) and vertical (Makino et al., 2011; Martin et al., 2012) transmission of infectious agents. Interestingly, new epidemic models have been recently developed that make use of MLST data to infer social network structure in ubiquitous commensal bacteria too (Plucinski et al., 2011). MLST has also helped to investigate the emergence and spread of antibiotic resistance to meticillin, erythromycin, macrolides and quinolones (Atkinson et al., 2009; De Francesco et al., 2011; Egger et al., 2012; Haran et al., 2012; Ibarz-Pavon et al., 2011; Pérez-Losada et al., 2007b; Tazi et al., 2010) and virulence (including virulent factors and genes and diseases associations) (Ch'ng et al., 2011; Dingle et al., 2011; Matsunari et al., 2012; Schultsz et al., 2012; Springman et al., 2009). It has also been used to monitor the effects of vaccination programs (pre and post-vaccine) (Adetifa et al., 2012; Climent et al., 2010; Hanage et al., 2011; Maiden and Stuart, 2002; Pichon et al., 2009; Stefanelli et al., 2009), improve vaccination strategies (Hanage et al., 2011; Racloz and Luiz, 2010; Stefanelli et al., 2009), and design new vaccines and new approaches to vaccination against *Streptococcus pneumoniae and N. meningitides* (Bambini et al., 2009; Pizza et al., 2000; Urwin et al., 2004). Finally, MLST has also contributed to the identification of sources of human infection from natural hosts (e.g., livestock animals and dogs) and environmental (e.g., animal-derived food) reservoirs (Bessell et al., 2012; Gripp et al., 2011; Ngo et al., 2011; O'Mahony et al., 2011), to identify host or niche associations (Hotchkiss et al., 2011; Sheppard et al., 2010a; Sproston et al., 2011) and zoonotic transmissions (Sahin et al., 2012; Sakwinska et al., 2011; Walther et al., 2012), and to study biological interactions like symbiosis in *Wolbachia* from insects (Russell et al., 2009).

4.2. Phylogenetics, taxonomy, and speciation

MLST data have been used to infer clone and species relationships and phylogroups in pathogenic (e.g., *Actinomyces*) and beneficial (e.g., *Oenococcus oeni* and *Trypanosoma cruzi*) microbiota (Bilhere et al., 2009; Bridier et al., 2010; Henssge et al., 2011; Yeo et al., 2011), separate and validate similar or sibling species of *Streptococcus oralis* and *Lactobacillus delbrueckii* (Do et al., 2009; Tanigawa and Watanabe, 2011) and identify new ones in, for example, the genera *Bartonella*, *Bacillus* and *Burkholderia* (Chaloner et al., 2011; Guinebretiere et al., 2012; Vanlaere et al., 2008, 2009), suggest new taxonomic classifications (e.g., *Lactococcus lactis*) (Passerini et al., 2010), validate COI barcodes in *Wolbachia* (Smith et al., 2012), and to discuss the bacterial species concept (Godreuil et al., 2005; Vos, 2011). MLST data are particularly useful for species diagnosis, as they provide both genealogical information as well as information on recombination (see below), which is critical for bacterial species identification (Dykhuizen and Green, 1991; Fraser et al., 2007), as revealed in *Streptococcus* (Ahmad et al., 2009).

4.3. Population structure and dynamics

MLST has been instrumental at confirming the clonal structure of many organisms like *S. aureus* (see Pérez-Losada et al., 2006 for a review); but also at identifying epidemic clonal complexes in other taxa like *Staphylococcus haemolyticus* (Cavanagh et al., 2012), *Yer-*

sinia pseudotuberculosis (Ch'ng et al., 2011) or *Streptococcus suis* (Schultsz et al., 2012); or even taxa considered non-clonal, such as *Pseudomonas aeruginosa* (Maatallah et al., 2011) or *Burkholderia pseudomallei* (Dale et al., 2011).

MLST data have been used to infer population structure at both temporal (de Filippis et al., 2012; Pérez-Losada et al., 2007c; Sproston et al., 2011) and geographical scales (Jorgensen et al., 2011) in for example *Neisseria* and *Campylobacter*, and to infer the epidemiological processes that may be responsible for the contemporary geographic distributions of diseases (phylogeography). For example, phylogeographic structure driven by host immunity has been detected in *S. aureus* from West China (Fan et al., 2009), while human activity has driven differentiation in *Clostridium difficile* isolates from North America, Europe, and Australia (Stabler et al., 2012). Similar studies based on MLST data have determined the geographic origin of *Mannheimia haemolytica* in European cattle and sheep (Petersen et al., 2009).

Another major contribution of MLST to bacterial population genetics has been the assessment of the relative impact of recombination and point mutation (the r/m ratio) in bacteria and archaea (Vos and Didelot, 2009) and within and among clones of, for example, *N. meningitidis*, *S. aureus*, *Y. pseudotuberculosis* or *Streptococcus dysgalactiae* (Basic-Hammer et al., 2010; Ch'ng et al., 2011; Feil et al., 1999, 2000; McMillan et al., 2010, 2011) or among species of *Streptococcus* (Ahmad et al., 2009; Do et al., 2010). MLST has also effectively identified the impact of selection in *Orientia tsutsugamushi*, *N. meningitidis*, *Bacillus cereus*, Group B Streptococcus or *Vibrio parahaemolyticus* (Duong et al., in press; Jolley et al., 2005; Raymond et al., 2010; Springman et al., 2009; Yan et al., 2011) and the contributors to population genetic diversity (see also Pérez-Losada et al., 2006). Similarly, MLST has provided insights on past population dynamics (epidemiological history), inferred as the variation in relative genetic diversity (or population size) since some time in the past, usually the time of emergence of the disease, in *Neisseria gonorrhoeae* (Pérez-Losada et al., 2007b,c; Tazi et al., 2010).

4.4. Other applications

MLST data have also been applied to biosafety research such as the detection of contamination with *S. aureus* in Portuguese public buses (Simoes et al., 2010), US West Coast public marine beaches (Soge et al., 2009), and in the working environment of many Swiss microbial laboratories (Schmidlin et al., 2010). Besides farm animals (above), MLST has also been applied in plant agriculture to identify genomospecies of *Pseudomonas syringae* causing bacterial leaf spot on parsley (Bull et al., 2011) and assess nodule occupancy of soybean by in *Bradyrhizobium* (Van Berkum et al., 2012), or to study the evolution of agriculture-associated disease caused by *Campylobacter coli* in farm animals from Scotland (Sheppard et al., 2010b). Another interesting application has been the tracing of ancient human migrations worldwide (Falush et al., 2003) or across India (Devi et al., 2007) and Malaysia (Tay et al., 2009), using *H. pylori* MLST data from human gastric mucosa.

Overall, MLST studies have both increased our knowledge of the diversity, population structure and dynamics of bacterial pathogens worldwide (basic research) and helped to design better strategies of control and treatment of the diseases caused by those pathogens (applied research), which ultimately has contributed to improve public health.

5. MLST in the genomic era

With advances in DNA sequencing technologies comes the natural question of whether or not MLST will continue to have utility in the next-gen $1000 human genome era. The great advantage of MLST is the unlinked survey of genetic variation at the DNA sequence level at a relatively cheap and efficient cost (Okoro et al., 2012). Yet the next-gen sequencing technologies are rapidly making these advantages mute (Chan et al., 2012). NGS also relieves some of the disadvantages of MLST (detailed above), including the need to have a genome of the target organism to begin with, the lack of broad application of individual loci across a diversity of species [because levels of genetic diversity and amounts of recombination vary across species for the same locus; but see Jolley et al. (2012a)], and shorter read lengths to avoid complications of recombination. Next we highlight two approaches for incorporating NGS into pathogen typing, first through single nucleotide polymorphism (SNP) analysis and second through whole genome sequence analysis. We then consider the bioinformatic implications and complications of dealing with this totally different volume of data and the associated challenges.

5.1. SNP discovery and typing

The first typing approach taking full advantage of whole-genome sequence data is that of SNP analysis. The central idea here is to get not just a single reference genome, as is the case with MLST typing, but a number of reference genomes to identify polymorphic sites within the genome. These sites or SNPs can then be used as diagnostic markers for specific species and/or strains within species, depending on the extent of variation in the species. Ideally, for species diagnostics based on SNPs, one is looking for fixed differences between species. Thus, the method becomes problematic if only a few reference genomes are used to establish whether variants are fixed or not within a species. This problem becomes worse when trying to diagnose strains within species, as many more samples are needed to effectively determine fixation of SNPs within strain and differences among strains. However, the advantage of SNPs is that they can provide broader genomic representation with less linkage (thereby lessening the potential impact of recombination). They are also relatively evolutionarily stable. Because these are genotypic data with character state information, they are amenable to robust phylogenetic and population genetic analyses (detailed above). SNP analyses have been used in pathogen population genetics for a number of years now with highly effective results (e.g., Filliol et al., 2006). Initially, SNPs were relatively expensive characters to develop for species typing; however, they have become highly efficient and effective for a variety of species. For example, Holt et al. (2010) used a survey of 2000 SNPs to identify strains of *Salmonella enterica* serovar Typhi causing a typhoid outbreak in children from Kathmandu, Nepal. More recently, Harris et al. (2012) used genome-wide SNPs of diverse *Chlamydia trachomatis* strains to identify phylogenetic relationships masked by recombination in current clinical typing (*ompA*). This study, hence, demonstrates how the whole genome data allow for the identification and therefore accommodation of recombination within the dataset and subsequent phylogenetic analyses.

5.2. Whole genome sequence typing (WGST)

With costs of whole genome sequencing coming down significantly through new technologies and better software (Ribeiro et al., 2012) and the need for whole genome data for both MLST and SNP approaches, recent studies have simply turned to eliminating these subsequent approaches for typing and used the whole genome data *per se*. The advantages of whole genome sequence typing (WGST) are clear – the highest resolution of genealogical data possible. This resolution has been instrumental to examine and reclassify species of *Neisseria* (Bennett et al., 2012). Here the authors studied the taxonomic relationships of 55 *Neisseria* representatives using 246 core genes (including 53 *rps* genes) and BIGS_{DB}. Variation in these genes identified seven species groups, which were not completely congruent with current species and isolate designations. Moreover, the seven groups could be reliably and rapidly identified using the *rps* genes, further confirming the efficiency and power of rMLST (as also demonstrated by Jolley et al., 2012a). Demonstrating the resolving power of WGST against other genetic (SNPs) and phenotypic (RFLP, VNTR) approaches in distinguishing strains of *Mycobacterium tuberculosis*, Schürch and van Soolingen (2012) argued that WGST will become the sole diagnostic tool of tuberculosis, including genetic characterization and drug resistance and susceptibility testing. However, others argued for a more integrated approach (combining SNP analysis with WGST), especially while sequencing costs are still high and may subject studies to issues of sampling bias (Pearson et al., 2009). But with technological advances occurring regularly, we are quickly moving to the full capacity of WGST (see Fig. 1 – WGS) as a standard operating procedure (Vogel et al., 2012). Studies have also shown that WGST and comparative genomics can reveal unique genetic elements missed by lesser resolution approaches such as SNP and MLST typing (Köser et al., 2012).

5.3. Bioinformatic considerations

Despite the significant promise of next generation sequencing techniques leading to whole genome sequence typing for pathogens, the move to whole genome analysis is not without challenges. The most significant of these is the ability to analyze this new volume of data in a reasonable and efficient manner. In this regard, Jolley et al. (2012b) demonstrate how whole-genome data from a meningococcal disease outbreak can be analyzed in real time by investigators using the analytical tools integrated into the PubMLST.org website.

With WGST comes also the need for genome assembly which can be fraught with difficulty (Schatz et al., 2010) and thereby introduce errors in assembled genomes that will appear as strain specific variation. Thus, ultimate care must be taken with analyses of whole genome data both at the assembly stage and downstream analyses. One approach to deal with this volume of data is to relate these whole genome sequence data back to MLST (Larsen et al., 2012). However, this approach then looses the advantages of WGST over MLST, including a broader survey of genetic signatures that are often critical in identifying causal agents of pathogenic outbreaks (e.g., Eppinger et al., 2011). An alternative approach is to map raw sequence reads to a reference database of pathogens for rapid and efficient identification of pathogens associated with a next-gen sequencing run from a biological sample (Clement et al., 2010). This approach has the advantage of avoiding the assembly step altogether, but requires a robust reference library of genomes to query against. No doubt substantial methodological advances will occur as more and more whole genome sequence data sets become available for consideration (e.g., Ribeiro et al., 2012).

6. Conclusions and prospects

MLST has played a major role in diagnosing pathogens of human disease. Rapid identification of such pathogens is crucial in our ability to identify, track, and treat disease outbreaks. MLST has proven to be a high-resolution genetic approach that provides data amenable to sophisticated phylogenetic and population genetic analyses. However, with the decrease in cost of genome sequencing, researchers are already moving to whole genome sequence analyses for such studies. We are clearly in the transition phase moving from MLST to whole genome sequencing typing and this shift provides extensive opportunity for the development of novel methodologies to accommodate this increased volume of genomic information.

Acknowledgements

We thank Michel Tibayrenc for inviting us to write this review. We were supported by CIBIO (Portugal) and George Washington University. P.C. was funded by a fellowship from Fundación Caja Madrid, Spain. E.C.N. was funded by Comisión Nacional de Investigación Científica y Tecnológica (CONICYT), Gobierno de Chile, Becas, and by BYU Graduate Mentoring (2011) and Research (2012) Awards.

References

Aanensen, D.M., Huntley, D.M., Feil, E.J., Spratt, B.G., 2009. EpiCollect: linking smartphones to web applications for epidemiology, ecology and community data collection. PLoS ONE 4, e6968.

Acinas, S.G., Klepac-Ceraj, V., Hunt, D.E., Pharino, C., Ceraj, I., Distel, D.L., Polz, M.F., 2004. Fine-scale phylogenetic architecture of a complex bacterial community. Nature 430, 551–554.

Adetifa, I.M., Antonio, M., Okoromah, C.A., Ebruke, C., Inem, V., Nsekpong, D., Bojang, A., Adegbola, R.A., 2012. Pre-vaccination nasopharyngeal pneumococcal carriage in a Nigerian population: epidemiology and population biology. PLoS ONE 7, e30548.

Ahmad, Y., Gertz Jr., R.E., Li, Z., Sakota, V., Broyles, L.N., Van Beneden, C., Facklam, R., Shewmaker, P.L., Reingold, A., Farley, M.M., Beall, B.W., 2009. Genetic relationships deduced from emm and multilocus sequence typing of invasive *Streptococcus dysgalactiae subsp. equisimilis* and *S. canis* recovered from isolates collected in the United States. J. Clin. Microbiol. 47, 2046–2054.

Atkinson, S.R., Paul, J., Sloan, E., Curtis, S., Miller, R., 2009. The emergence of meticillin-resistant *Staphylococcus aureus* among injecting drug users. J. Infect. 58, 339–345.

Auton, A., McVean, G., 2012. Estimating Recombination Rates from Genetic Variation in Humans, in: Anisimova, M. (Ed.), Evolutionary Genomics. Humana Press, pp. 217–237.

Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M.A., Alekseyenko, A.V., 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol. Biol. Evol. 29, 2157–2167.

Baker, S., Hanage, W.P., Holt, K.E., 2010. Navigating the future of bacterial molecular epidemiology. Curr. Opin. Microbiol. 13, 640–645.

Bambini, S., Muzzi, A., Olcen, P., Rappuoli, R., Pizza, M., Comanducci, M., 2009. Distribution and genetic variability of three vaccine components in a panel of strains representative of the diversity of serogroup B *meningococcus*. Vaccine 27, 2794–2803.

Bashalkhanov, S., Pandey, M., Rajora, O., 2009. A simple method for estimating genetic diversity in large populations from finite sample sizes. BMC Genet. 10, 84.

Basic-Hammer, N., Vogel, V., Basset, P., Blanc, D.S., 2010. Impact of recombination on genetic variability within *Staphylococcus aureus* clonal complexes. Infect. Genet. Evol. 10, 1117–1123.

Bennett, J.S., Jolley, K.A., Earle, S.G., Corton, C., Bentley, S.D., Parkhill, J., Maiden, M.C.J., 2012. A genomic approach to bacterial taxonomy: an examination and proposed reclassification of species within the genus *Neisseria*. Microbiology 158, 1570–1580.

Bessell, P.R., Rotariu, O., Innocent, G.T., Smith-Palmer, A., Strachan, N.J., Forbes, K.J., Cowden, J.M., Reid, S.W., Matthews, L., 2012. Using sequence data to identify alternative routes and risk of infection: a case-study of *Campylobacter* in Scotland. BMC Infect. Dis. 12, 80.

Bielejec, F., Rambaut, A., Suchard, M.A., Lemey, P., 2011. SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics. Bioinformatics 27, 2910–2912.

Bilhere, E., Lucas, P.M., Claisse, O., Lonvaud-Funel, A., 2009. Multilocus sequence typing of *Oenococcus oeni*: detection of two subpopulations shaped by intergenic recombination. Appl. Environ. Microbiol. 75, 1291–1300.

Boers, S.A., van der Reijden, W.A., Jansen, R., 2012. High-throughput multilocus sequence typing: bringing molecular typing to the next level. PLoS ONE 7, e39630.

Brehony, C., Jolley, K.A., Maiden, M.C., 2007. Multilocus sequence typing for global surveillance of meningococcal disease. FEMS Microbiol. Rev. 31, 15–26.

Bridier, J., Claisse, O., Coton, M., Coton, E., Lonvaud-Funel, A., 2010. Evidence of distinct populations and specific subpopulations within the species *Oenococcus oeni*. Appl. Environ. Microbiol. 76, 7754–7764.

Bull, C.T., Clarke, C.R., Cai, R., Vinatzer, B.A., Jardini, T.M., Koike, S.T., 2011. Multilocus sequence typing of *Pseudomonas syringae* sensu lato confirms previously described genomospecies and permits rapid identification of *P. syringae pv. coriandricola* and *P. syringae pv. apii* causing bacterial leaf spot on parsley. Phytopathology 101, 847–858.

Bybee, S.M., Bracken-Grissom, H., Haynes, B.D., Hermansen, R.A., Byers, R.L., Clement, M.J., Udall, J.A., Wilcox, E.R., Crandall, K.A., 2011. Targeted amplicon sequencing (TAS): a scalable next-gen approach to multilocus, multitaxa phylogenetics. Genome Biol. Evol. 3, 1312–1323.

Byrnes, E.J., Li, W., Lewit, Y., Ma, H., Voelz, K., Ren, P., Carter, D.A., Chaturvedi, V., Bildfell, R.J., May, R.C., Heitman, J., 2010. Emergence and pathogenicity of highly virulent *Cryptococcus gattii* genotypes in the northwest United States. PLoS Pathog. 6, e1000850.

Cassens, I., Mardulyn, P., Milinkovitch, M.C., 2005. Evaluating intraspecific "network" construction methods using simulated sequence data: do existing algorithms outperform the global maximum parsimony approach? Syst. Biol. 54, 363–372.

Castro-Nallar, E., Crandall, K.A., Pérez-Losada, M., 2012. Genetic diversity and molecular epidemiology of HIV transmission. Future Virol. 7, 239–252.

Cavanagh, J.P., Klingenberg, C., Hanssen, A.M., Fredheim, E.A., Francois, P., Schrenzel, J., Flaegstad, T., Sollid, J.E., 2012. Core genome conservation of *Staphylococcus haemolyticus* limits sequence based population structure analysis. J. Microbiol. Methods 89, 159–166.

Chaloner, G.L., Palmira, V., Birtles, R.J., 2011. Multi-locus sequence analysis reveals profound genetic diversity among isolates of the human pathogen *Bartonella bacilliformis*. PLoS Negl. Trop. Dis. 5, e1248.

Chan, M.S., Maiden, M.C.J., Spratt, B.G., 2001. Database-driven multi locus sequence typing (MLST) of bacterial pathogens. Bioinformatics 17, 1077–1083.

Chan, J.Z.-M., Pallen, M.J., Oppenheim, B., Constantinidou, C., 2012. Genome sequencing in clinical microbiology. Nature Biotechnol 30, 1068–1071.

Cheng, L., Connor, T.R., Aanensen, D.M., Spratt, B.G., Corander, J., 2011. Bayesian semi-supervised classification of bacterial samples using MLST databases. Bioinformatics 12, 302.

Ch'ng, S.L., Octavia, S., Xia, Q., Duong, A., Tanaka, M.M., Fukushima, H., Lan, R., 2011. Population structure and evolution of pathogenicity of *Yersinia pseudotuberculosis*. Appl. Environ. Microbiol. 77, 768–775.

Clement, N.L., Snell, Q., Clement, M.J., Hollenhorst, P.C., Purwar, J., Graves, B.J., Cairns, B.R., Johnson, W.E., 2010. The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. Bioinformatics 26, 38–45.

Climent, Y., Urwin, R., Yero, D., Martinez, I., Martin, A., Sotolongo, F., Maiden, M.C., Pajon, R., 2010. The genetic structure of *Neisseria meningitidis* populations in Cuba before and after the introduction of a serogroup BC vaccine. Infect. Genet. Evol. 10, 546–554.

Comas, I., Homolka, S., Niemann, S., Gagneux, S., 2009. Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. PLoS ONE 4, e7815.

Cooper, J.E., Feil, E.J., 2006. The phylogeny of *Staphylococcus aureus* – which genes make the best intra-species markers? Microbiology 152, 1297–1305.

Crandall, K.A., Pérez-Losada, M., 2008. Epidemiological and evolutionary dynamics of pathogens. In: Baquero, F., Nombela, C., Cassell, G.H., Gutiérrez-Fuentes, J.A. (Eds.), Evolutionary Biology of Bacterial and Fungal Pathogens. ASM Press, Washington, DC, pp. 21–30.

Cronn, R., Knaus, B.J., Liston, A., Maughan, P.J., Parks, M., Syring, J.V., Udall, J., 2012. Targeted enrichment strategies for next-generation plant biology. Am. J. Bot. 99, 291–311.

Dale, J., Price, E.P., Hornstra, H., Busch, J.D., Mayo, M., Godoy, D., Wuthiekanun, V., Baker, A., Foster, J.T., Wagner, D.M., Tuanyok, A., Warner, J., Spratt, B.G., Peacock, S.J., Currie, B.J., Keim, P., Pearson, T., 2011. Epidemiological tracking and population assignment of the non-clonal bacterium, *Burkholderia pseudomallei*. PLoS Negl. Trop. Dis. 5, e1381.

Darling, A.E., Mau, B., Perna, N.T., 2010. ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS ONE 5, e11147.

Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2012. JModelTest 2: more models, new heuristics and parallel computing. Nat. Methods 9, 772.

de Filippis, I., de Lemos, A.P., Hostetler, J.B., Wollenberg, K., Sacchi, C.T., Harrison, L.H., Bash, M.C., Prevots, D.R., 2012. Molecular epidemiology of *Neisseria meningitidis* serogroup B in Brazil. PLoS ONE 7, e33016.

De Francesco, M.A., Caracciolo, S., Gargiulo, F., Manca, N., 2011. Phenotypes, genotypes, serotypes and molecular epidemiology of erythromycin-resistant *Streptococcus agalactiae* in Italy. Eur. J. Clin. Microbiol. Infect. Dis. 31, 1741–1747.

Devi, S.M., Ahmed, I., Francalacci, P., Hussain, M.A., Akhter, Y., Alvi, A., Sechi, L.A., Megraud, F., Ahmed, N., 2007. Ancestral European roots of *Helicobacter pylori* in India. BMC Genomics 8, 184.

Didelot, X., Falush, D., 2007. Inference of bacterial microevolution using multilocus sequence data. Genetics 175, 1251–1266.

Didelot, X., Maiden, M.C.J., 2010. Impact of recombination on bacterial evolution. Trends Microbiol. 18, 315–322.

Dingle, K.E., Griffiths, D., Didelot, X., Evans, J., Vaughan, A., Kachrimanidou, M., Stoesser, N., Jolley, K.A., Golubchik, T., Harding, R.M., Peto, T.E., Fawley, W., Walker, A.S., Wilcox, M., Crook, D.W., 2011. Clinical *Clostridium difficile*: clonality and pathogenicity locus diversity. PLoS ONE 6, e19993.

Do, T., Jolley, K.A., Maiden, M.C., Gilbert, S.C., Clark, D., Wade, W.G., Beighton, D., 2009. Population structure of *Streptococcus oralis*. Microbiology 155, 2593–2602.

Do, T., Gilbert, S.C., Clark, D., Ali, F., Fatturi Parolo, C.C., Maltz, M., Russell, R.R., Holbrook, P., Wade, W.G., Beighton, D., 2010. Generation of diversity in *Streptococcus mutans* genes demonstrated by MLST. PLoS ONE 5, e9073.

Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214.

Drummond, A.J., Rambaut, A., Shapiro, B., Pybus, O.G., 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. 22, 1185–1192.

DuBose, R.F., Dykhuizen, D.E., Hartl, D.L., 1988. Genetic exchange among natural isolates of bacteria: Recombination within the *phoA* gene of *Escherichia coli*. Proc. Natl. Acad. Sci. U.S.A. 85, 7036–7040.

Duong, V., Blassdell, K., May, T.T., Sreyrath, L., Gavotte, L., Morand, S., Frutos, R., Buchy, P., in press. Diversity of *Orientia tsutsugamushi* clinical isolates in Cambodia reveals active selection and recombination process. Infect. Genet. Evol.

Dykhuizen, D.E., Green, L., 1991. Recombination in *Escherichia coli* and the definition of biological species. J. Bacteriol. 173, 7257–7268.

Egger, R., Korczak, B.M., Niederer, L., Overesch, G., Kuhnert, P., 2012. Genotypes and antibiotic resistance of *Campylobacter coli* in fattening pigs. Vet. Microbiol. 155, 272–278.

Elberse, K.E.M., Nunes, S., Sá-Leão, R., van der Heide, H.G.J., Schouls, L.M., 2011. Multiple-locus variable number tandem repeat analysis for *Streptococcus pneumoniae*: comparison with PFGE and MLST. PLoS ONE 6, e19668.

Enright, M.C., Spratt, B.G., 1998. A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. Microbiology 144, 3049–3060.

Eppinger, M., Mammel, M.K., Leclerc, J.E., Ravel, J., Cebula, T.A., 2011. Genomic anatomy of *Escherichia coli* O157:H7 outbreaks. Proc. Natl. Acad. Sci. U.S.A. 108, 20142–20147.

Erali, M., Voelkerding, K.V., Wittwer, C.T., 2008. High resolution melting applications for clinical laboratory medicine. Exp. Mol. Pathol. 85, 50–58.

Ewing, B., Green, P., 1998. Base-calling of automated sequencer traces using phred II. Error probabilities. Genome Res. 8, 186–194.

Ewing, B., Hillier, L., Wendl, M.C., Green, P., 1998. Base-calling of automated sequencer traces using phred I. Accuracy assessment. Genome Res. 8, 175–185.

Excoffier, L., Heckel, G., 2006. Computer programs for population genetics data analysis: a survival guide. Nat. Rev. Genet. 7, 745–758.

Faircloth, B.C., McCormack, J.E., Crawford, N.G., Harvey, M.G., Brumfield, R.T., Glenn, T.C., 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61, 717–726.

Falush, D., Wirth, T., Linz, B., Pritchard, J.K., Stephens, M., Kidd, M., Blaser, M.J., Graham, D.Y., Vacher, S., Pérez-Pérez, G.I., Yamaoka, Y., Mégraud, F., Otto, K., Reichard, U., Katzowitsch, E., Wang, X., Achtman, M., Suerbaum, S., 2003. Traces of human migrations in *Helicobacter pylori* populations. Science 299, 1582–1585.

Fan, J., Shu, M., Zhang, G., Zhou, W., Jiang, Y., Zhu, Y., Chen, G., Peacock, S.J., Wan, C., Pan, W., Feil, E.J., 2009. Biogeography and virulence of *Staphylococcus aureus*. PLoS ONE 4, e6216.

Feil, E.J., Maiden, M.C.J., Achtman, M., Spratt, B.G., 1999. The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. Mol. Biol. Evol. 16, 1496–1502.

Feil, E.J., Enright, M.C., Spratt, B.G., 2000. Estimating the relative contributions of mutation and recombination to clonal diversification: a comparison between *Neisseria meningitidis* and *Streptococcus pneumoniae*. Res. Microbiol. 151, 465–469.

Feil, E.J., Li, B.C., Aanensen, D.M., Hanage, W.P., Spratt, B.G., 2004. EBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J. Bacteriol. 186, 1518–1530.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17, 368–376.

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.

Ferreira, R., Borges, V., Nunes, A., Nogueira, P.J., Borrego, M.J., Gomes, J.P., 2012. Impact of loci nature on estimating recombination and mutation rates in *Chlamydia trachomatis*. Genes Genomes Genet. 2, 761–768.

Filliol, I., Motiwala, A.S., Cavatore, M., Qi, W., Hazbon, M.H., Bobadilla del Valle, M., Fyfe, J., Garcia-Garcia, L., Rastogi, N., Sola, C., Zozio, T., Guerrero, M.I., Leon, C.I., Crabtree, J., Angiuoli, S., Eisenach, K.D., Durmaz, R., Joloba, M.L., Rendon, A., Sifuentes-Osornio, J., Ponce de Leon, A., Cave, M.D., Fleischmann, R., Whittam, T.S., Alland, D., 2006. Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. J. Bacteriol. 188, 759–772.

Foxman, B., Zhang, L., Koopman, J., Manning, S., Marrs, C., 2005. Choosing an appropriate bacterial typing technique for epidemiologic studies. Epidemiologic Perspectives & Innovations 2, 10.

Francisco, A.P., Bugalho, M., Ramirez, M., Carrico, J.A., 2009. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. Bioinformatics 10, 152.

Francisco, A.P., Vaz, C., Monteiro, P.T., Melo-Cristino, J., Ramirez, M., Carrio, J.A., 2012. PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. BMC Bioinformatics 13, 87.

Fraser, C., Hanage, W.P., Spratt, B.G., 2007. Recombination and the nature of bacterial speciation. Science 315, 476–480.

Godreuil, S., Cohan, F., Shah, H., Tibayrenc, M., 2005. Which species concept for pathogenic bacteria? An E-Debate. Infect. Genet. Evol. 5, 375–387.

Gray, R.R., Tatem, A.J., Johnson, J.A., Alekseyenko, A.V., Pybus, O.G., Suchard, M.A., Salemi, M., 2011. Testing Spatiotemporal Hypothesis of Bacterial Evolution Using Methicillin-Resistant *Staphylococcus aureus* ST239 Genome-wide Data within a Bayesian Framework. Mol. Biol. Evol. 28, 1593–1603.

Gripp, E., Hlahla, D., Didelot, X., Kops, F., Maurischat, S., Tedin, K., Alter, T., Ellerbroek, L., Schreiber, K., Schomburg, D., Janssen, T., Bartholomaus, P., Hofreuter, D., Woltemate, S., Uhr, M., Brenneke, B., Gruning, P., Gerlach, G., Wieler, L., Suerbaum, S., Josenhans, C., 2011. Closely related *Campylobacter* *jejuni* strains from different sources reveal a generalist rather than a specialist lifestyle. Genomics 12, 584.

Grundmann, H., Aanensen, D.M., Van Den Wijngaard, C.C., Spratt, B.G., Harmsen, D., Friedrich, A.W., 2010. Geographic distribution of *Staphylococcus aureus* causing invasive infections in Europe: a molecular-epidemiological analysis. PLoS Med. 7, e1000215.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321.

Guinebretiere, M.H., Auger, S., Galleron, N., Contzen, M., De Sarrau, B., De Buyser, M.L., Lamberet, G., Fagerlund, A., Granum, P.E., Lereclus, D., De Vos, P., Nguyen-The, C., Sorokin, A., 2012. *Bacillus cytotoxicus* sp. nov. is a new thermotolerant species of the *Bacillus cereus* group occasionally associated with food poisoning. Int. J. Syst. Evol. Microbiol.. http://dx.doi.org/10.1099/ijs.0.030627-0.

Hall, B.G., Barlow, M., 2006. Phylogenetic analysis as a tool in molecular epidemiology of infectious diseases. Ann. Epidemiol. 16, 157–169.

Hall, B.G., Ehrlich, G.D., Hu, F.Z., 2010. Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. Microbiology 156, 1060–1068.

Hanage, W.P., Bishop, C.J., Lee, G.M., Lipsitch, M., Stevenson, A., Rifas-Shiman, S.L., Pelton, S.I., Huang, S.S., Finkelstein, J.A., 2011. Clonal replacement among 19A *Streptococcus pneumoniae* in Massachusetts, prior to 13 valent conjugate vaccination. Vaccine 29, 8877–8881.

Haran, K.P., Godden, S.M., Boxrud, D., Jawahir, S., Bender, J.B., Sreevatsan, S., 2012. Prevalence and characterization of *Staphylococcus aureus*, including methicillin-resistant *Staphylococcus aureus,* isolated from bulk tank milk from Minnesota dairy farms. J. Clin. Microbiol. 50, 688–695.

Harbottle, H., White, D., McDermott, P., Walker, R., Zhao, S., 2006. Comparison of multilocus sequence typing, pulsed-field gel electrophoresis, and antimicrobial susceptibility typing for characterization of *Salmonella enterica* serotype Newport isolates. J. Clin. Microbiol. 44, 2449–2457.

Harismendy, O., Ng, P., Strausberg, R., Wang, X., Stockwell, T., Beeson, K., Schork, N., Murray, S., Topol, E., Levy, S., Frazer, K., 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. 10, R32.

Harris, S.R., Clarke, I.N., Seth-Smith, H.M.B., Solomon, A.W., Cutcliffe, L.T., Marsh, P., Skilton, R.J., Holland, M.J., Mabey, D., Peeling, R.W., 2012. Whole-genome analysis of diverse *Chlamydia trachomatis* strains identifies phylogenetic relationships masked by current clinical typing. Nat. Genet. 44, 413–419.

Heled, J., Drummond, A.J., 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27, 570–580.

Henssge, U., Do, T., Gilbert, S.C., Cox, S., Clark, D., Wickstrom, C., Ligtenberg, A.J., Radford, D.R., Beighton, D., 2011. Application of MLST and pilus gene sequence comparisons to investigate the population structures of *Actinomyces naeslundii* and *Actinomyces oris*. PLoS ONE 6, e21430.

Heym, B., Le Moal, M., Armand-Lefevre, L., Nicolas-Chanoine, M.H., 2002. Multilocus sequence typing (MLST) shows that the 'Iberian'clone of methicillin-resistant *Staphylococcus aureus* has spread to France and acquired reduced susceptibility to teicoplanin. J. Antimicrob. Chemother. 50, 323–329.

Hird, S.M., Brumfield, R.T., Carstens, B.C., 2011. PRGmatic: an efficient pipeline for collating genome-enriched second-generation sequencing data using a 'provisional-reference genome'. Mol Ecol Resour 11, 743–748.

Hoff, K., 2009. The effect of sequencing errors on metagenomic gene prediction. BMC Genomics 10, 520.

Holt, K.E., Baker, S., Dongol, S., Basnyat, B., Adhikari, N., Thorson, S., Pulickal, A.S., Song, Y., Parkhill, J., Farrar, J.J., Murdoch, D.R., Kelly, D.F., Pollard, A.J., Dougan, G., 2010. High-throughput bacterial SNP typing identifies distinct clusters of *Salmonella typhi* causing typhoid in Nepalese children. BMC Infect. Dis. 10, 144.

Hotchkiss, E.J., Hodgson, J.C., Lainson, F.A., Zadoks, R.N., 2011. Multilocus sequence typing of a global collection of *Pasteurella multocida* isolates from cattle and other host species demonstrates niche association. BMC Microbiol. 11, 115.

Huelsenbeck, J.P., Ronquist, F., Nielsen, R., Bollback, J.P., 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294, 2310–2314.

Hui, P., 2012. Next generation sequencing: chemistry, technology and applications. Top. Curr. Chem., 1–18.

Huson, D.H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23, 254–267.

Ibarz-Pavon, A.B., Morais, L., Sigauque, B., Mandomando, I., Bassat, Q., Nhacolo, A., Quinto, L., Soriano-Gabarro, M., Alonso, P.L., Roca, A., 2011. Epidemiology, molecular characterization and antibiotic resistance of *Neisseria meningitidis* from patients ⩽ 15 years in Manhica, rural Mozambique. PLoS ONE 6, e19717.

Jeffrey, D.W., 2004. Estimating recombination rates using three-site likelihoods. Genetics 167, 1461–1473.

Jolley, K.A., 2009. Internet-based sequence-typing databases for bacterial molecular epidemiology. Methods Mol. Evol. 551, 305–312.

Jolley, K.A., Maiden, M.C.J., 2006. AgdbNet–antigen sequence database software for bacterial typing. BMC Bioinformatics 7, 314.

Jolley, K.A., Maiden, M.C.J., 2010. BIGSdb: scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11, 595.

Jolley, K.A., Feil, E.J., Chan, M.S., Maiden, M.C., 2001. Sequence type analysis and recombinational tests (START). Bioinformatics 17, 1230–1231.

Jolley, K.A., Chan, M.S., Maiden, M.C.J., 2004. MlstdbNet–distributed multi-locus sequence typing (MLST) databases. BMC Bioinformatics 5, 86.

Jolley, K.A., Wilson, D.J., Kriz, P., McVean, G., Maiden, M.C., 2005. The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in *Neisseria meningitidis*. Mol. Biol. Evol. 22, 562–569.

Jolley, K.A., Bliss, C.M., Bennett, J.S., Bratcher, H.B., Brehony, C.M., Colles, F.M., Wimalarathna, H.M., Harrison, O.B., Sheppard, S.K., Cody, A.J., 2012a. Ribosomal multi-locus sequence typing: universal characterisation of bacteria from domain to strain. Microbiology 158, 1005–1015.

Jolley, K.A., Hill, D.M.C., Bratcher, H.B., Harrison, O.B., Feavers, I.M., Parkhill, J., Maiden, M.C.J., 2012b. Resolution of a meningococcal disease outbreak from whole-genome sequence data with rapid web-based analysis methods. J. Clin. Microbiol. 50, 3046–3053.

Jorgensen, F., Ellis-Iversen, J., Rushton, S., Bull, S.A., Harris, S.A., Bryan, S.J., Gonzalez, A., Humphrey, T.J., 2011. Influence of season and geography on *Campylobacter jejuni* and *C. coli* subtypes in housed broiler flocks reared in Great Britain. Appl. Environ. Microbiol. 77, 3741–3748.

Katoh, K., Kuma, K., Toh, H., Miyata, T., 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33, 511–518.

Kelsey, C.R., Crandall, K.A., Voevodin, A.F., 1999. Different models, different trees: the geographic origin of PTLV-I. Mol. Phylogenet. Evol. 13, 336–347.

Killgore, G., Thompson, A., Johnson, S., Brazier, J., Kuijper, E., Pepin, J., Frost, E.H., Savelkoul, P., Nicholson, B., Van Den Berg, R.J., 2008. Comparison of seven techniques for typing international epidemic strains of *Clostridium difficile*: restriction endonuclease analysis, pulsed-field gel electrophoresis, PCR-ribotyping, multilocus sequence typing, multilocus variable-number tandem-repeat analysis, amplified fragment length polymorphism, and surface layer protein A gene sequence typing. J. Clin. Microbiol. 46, 431–437.

Kosakovsky Pond, S.L., Frost, S.D.W., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21, 676–679.

Köser, C.U., Ellington, M.J., Cartwright, E.J.P., Gillespie, S.H., Brown, N.M., Farrington, M., Holden, M.T.G., Dougan, G., Bentley, S.D., Parkhill, J., 2012. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog. 8, e1002824.

Kriz, P., Kalmusova, J., Felsberg, J., 2002. Multilocus sequence typing of *Neisseria meningitidis* directly from cerebrospinal fluid. Epidemiol. Infect. 128, 157–160.

Kubatko, L.S., Carstens, B.C., Knowles, L.L., 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25, 971–973.

Kuhn, G., Francioli, P., Blanc, D., 2006. Evidence for clonal evolution among highly polymorphic genes in methicillin-resistant *Staphylococcus aureus*. J. Bacteriol. 188, 169–178.

Kuhner, M.K., 2006. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics 22, 768–770.

Larsen, M.V., Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R.L., Jelsbak, L., Sicheritz-Ponten, T., Ussery, D.W., Aarestrup, F.M., Lund, O., 2012. Multilocus sequence typing of total-genome-sequenced bacteria. J. Clin. Microbiol. 50, 1355–1361.

Lefebvre, J.F., Labuda, D., 2008. Fraction of informative recombinations: a heuristic approach to analyze recombination rates. Genetics 178, 2069–2079.

Lemey, P., Rambaut, A., Drummond, A.J., Suchard, M.A., 2009. Bayesian phylogeography finds its roots. PLoS Comput. Biol. 5, e1000520.

Lemey, P., Rambaut, A., Welch, J.J., Suchard, M.A., 2010. Phylogeography takes a relaxed random walk in continuous space and time. Mol. Biol. Evol. 27, 1877–1885.

Lemmon, A.R., Emme, S.A., Lemmon, E.M., 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61, 727–744.

Lewis-Rogers, N., Bendall, M.L., Crandall, K.A., 2009. Phylogenetic relationships and molecular adaptation dynamics of human rhinoviruses. Mol. Biol. Evol. 26, 969–981.

Li, W., Raoult, D., Fournier, P.-E., 2009. Bacterial strain typing in the genomic era. FEMS Microbiol. Rev. 33, 892–916.

Liu, L., 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24, 2542–2543.

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

Maatallah, M., Cheriaa, J., Backhrouf, A., Iversen, A., Grundmann, H., Do, T., Lanotte, P., Mastouri, M., Elghmati, M.S., Rojo, F., Mejdi, S., Giske, C.G., 2011. Population structure of *Pseudomonas aeruginosa* from five Mediterranean countries: evidence for frequent recombination and epidemic occurrence of CC235. PLoS ONE 6, e25617.

Maiden, M.C.J., 2006. Multilocus sequence typing of bacteria. Annu. Rev. Microbiol. 60, 561–588.

Maiden, M.C., Stuart, J.M., 2002. Carriage of serogroup C *meningococci* 1 year after meningococcal C conjugate polysaccharide vaccination. Lancet 359, 1829–1831.

Maiden, M.C.J., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. U.S.A. 95, 3140.

Makino, H., Kushiro, A., Ishikawa, E., Muylaert, D., Kubota, H., Sakai, T., Oishi, K., Martin, R., Ben Amor, K., Oozeer, R., Knol, J., Tanaka, R., 2011. Transmission of intestinal *Bifidobacterium longum* subsp. *longum* strains from mother to infant, determined by multilocus sequencing typing and amplified fragment length polymorphism. Appl. Environ. Microbiol. 77, 6788–6793.

Malachowa, N., Sabat, A., Gniadkowski, M., Krzyszton-Russjan, J., Empel, J., Miedzobrodzki, J., Kosowska-Shick, K., Appelbaum, P.C., Hryniewicz, W., 2005. Comparison of multiple-locus variable-number tandem-repeat analysis with pulsed-field gel electrophoresis, spa typing, and multilocus sequence typing for clonal characterization of *Staphylococcus aureus* isolates. J. Clin. Microbiol. 43, 3095–3100.

Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J., Turner, D.J., 2010. Target-enrichment strategies for next-generation sequencing. Nat. Methods 7, 111–118.

Martin, D.P., Lemey, P., Lott, M., Moulton, V., Posada, D., Lefeuvre, P., 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26, 2462–2463.

Martin, D.P., Lemey, P., Posada, D., 2011. Analysing recombination in nucleotide sequences. Mol Ecol Resour 11, 943–955.

Martin, V., Maldonado-Barragan, A., Moles, L., Rodriguez-Banos, M., Campo, R.D., Fernandez, L., Rodriguez, J.M., Jimenez, E., 2012. Sharing of bacterial strains between breast milk and infant feces. J. Hum. Lact. 28, 36–44.

Matsunari, O., Shiota, S., Suzuki, R., Watada, M., Kinjo, N., Murakami, K., Fujioka, T., Kinjo, F., Yamaoka, Y., 2012. Association between *Helicobacter pylori* virulence factors and gastroduodenal diseases in Okinawa, Japan. J. Clin. Microbiol. 50, 876–883.

McAdam, P.R., Templeton, K.E., Edwards, G.F., Holden, M.T.G., Feil, E.J., Aanensen, D.M., Bargawi, H.J.A., Spratt, B.G., Bentley, S.D., Parkhill, J., Enright, M.C., Holmes, A., Girvan, E.K., Godfrey, P.A., Feldgarden, M., Kearns, A.M., Rambaut, A., Robinson, D.A., Fitzgerald, J.R., 2012. Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant *Staphylococcus aureus*. Proc. Natl. Acad. Sci. U.S.A. 109, 9107–9112.

McMillan, D.J., Bessen, D.E., Pinho, M., Ford, C., Hall, G.S., Melo-Cristino, J., Ramirez, M., 2010. Population genetics of *Streptococcus dysgalactiae* subspecies *equisimilis* reveals widely dispersed clones and extensive recombination. PLoS ONE 5, e11741.

McMillan, D.J., Kaul, S.Y., Bramhachari, P.V., Smeesters, P.R., Vu, T., Karmarkar, M.G., Shaila, M.S., Sriprakash, K.S., 2011. Recombination drives genetic diversification of *Streptococcus dysgalactiae* subspecies *equisimilis* in a region of streptococcal endemicity. PLoS ONE 6, e21346.

Medini, D., Serruto, D., Parkhill, J., Relman, D.A., Donati, C., Moxon, R., Falkow, S., Rappuoli, R., 2008. Microbiology in the post-genomic era. Nat. Rev. Microbiol. 6, 419–430.

Melles, D.C., van Leeuwen, W.B., Snijders, S.V., Horst-Kreft, D., Peeters, J.K., Verbrugh, H.A., van Belkum, A., 2007. Comparison of multilocus sequence typing (MLST), pulsed-field gel electrophoresis (PFGE), and amplified fragment length polymorphism (AFLP) for genetic typing of *Staphylococcus aureus*. J. Microbiol. Methods 69, 371–375.

Metzker, M.L., 2010. Sequencing technologies - the next generation. Nat. Rev. Genet. 11, 31–46.

Millat, G., Chanavat, V., Julia, S., Crehalet, H., Bouvagnet, P., Rousson, R., 2009. Validation of high-resolution DNA melting analysis for mutation scanning of the LMNA gene. Clin. Biochem. 42, 892–898.

Minin, V.N., Bloomquist, E.W., Suchard, M.A., 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol. Biol. Evol. 25, 1459–1471.

Ngo, T.H., Tran, T.B., Tran, T.T., Nguyen, V.D., Campbell, J., Pham, H.A., Huynh, H.T., Nguyen, V.V., Bryant, J.E., Tran, T.H., Farrar, J., Schultsz, C., 2011. Slaughterhouse pigs are a major reservoir of *Streptococcus suis* serotype 2 capable of causing human infection in southern Vietnam. PLoS ONE 6, e17943.

Okoro, C.K., Kingsley, R.A., Connor, T.R., Harris, S.R., Parry, C.M., Al-Mashhadani, M.N., Kariuki, S., Msefula, C.L., Gordon, M.A., de Pinna, E., 2012. Intracontinental spread of human invasive *Salmonella typhimurium* pathovariants in sub-Saharan Africa. Nat. Genet. 44, 1215–1221.

O'Mahony, E., Buckley, J.F., Bolton, D., Whyte, P., Fanning, S., 2011. Molecular epidemiology of *Campylobacter* isolates from poultry production units in southern Ireland. PLoS ONE 6, e28490.

Padhukasahasram, B., Wall, J.D., Marjoram, P., Nordborg, M., 2006. Estimating recombination rates from single-nucleotide polymorphisms using summary statistics. Genetics 174, 1517–1528.

Palazzo, I.C., Pitondo-Silva, A., Levy, C.E., da Costa Darini, A.L., 2011. Changes in vancomycin-resistant *Enterococcus faecium* causing outbreaks in Brazil. J. Hosp. Infect. 79, 70–74.

Parkhill, J., Sebaihia, M., Preston, A., Murphy, L.D., Thomson, N., Harris, D.E., Holden, M.T.G., Churcher, C.M., Bentley, S.D., Mungall, K.L., 2003. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. Nat. Genet. 35, 32–40.

Passerini, D., Beltramo, C., Coddeville, M., Quentin, Y., Ritzenthaler, P., Daveran-Mingot, M.L., Le Bourgeois, P., 2010. Genes but not genomes reveal bacterial domestication of *Lactococcus lactis*. PLoS ONE 5, e15306.

Pearse, D.E., Crandall, K., 2004. Beyond Fst: analysis of population genetic data for conservation. Conserv. Genet. 5, 585–602.

Pearson, T., Okinaka, R.T., Foster, J.T., Keim, P., 2009. Phylogenetic understanding of clonal populations in an era of whole genome sequencing. Infect. Genet. Evol. 9, 1010–1019.

Pérez-Losada, M., Browne, E.B., Madsen, A., Wirth, T., Viscidi, R.P., Crandall, K.A., 2006. Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. Infect. Genet. Evol. 6, 97–112.

Pérez-Losada, M., Porter, M.L., Tazi, L., Crandall, K.A., 2007a. New methods for inferring population dynamics from microbial sequences. Infect. Genet. Evol. 7, 24–43.

Pérez-Losada, M., Crandall, K.A., Bash, M.C., Dan, M., Zenilman, J., Viscidi, R.P., 2007b. Distinguishing importation from diversification of quinolone-resistant *Neisseria gonorrhoeae* by molecular evolutionary analysis. BMC Evol. Biol. 7, 84.

Pérez-Losada, M., Crandall, K.A., Zenilman, J., Viscidi, R.P., 2007c. Temporal trends in gonococcal population genetics in a high prevalence urban community. Infect. Genet. Evol. 7, 271–278.

Pérez-Losada, M., Porter, M.L., Viscidi, R.P., Crandall, K.A., 2011. Multilocus sequence typing of pathogens. In: Tibayrenc, M. (Ed.), Genetics and Evolution of Infectious Diseases. Elsevier Inc., pp. 503–521.

Petersen, A., Christensen, H., Kodjo, A., Weiser, G.C., Bisgaard, M., 2009. Development of a multilocus sequence typing (MLST) scheme for *Mannheimia haemolytica* and assessment of the population structure of isolates obtained from cattle and sheep. Infect. Genet. Evol. 9, 626–632.

Pichon, B., Bennett, H.V., Efstratiou, A., Slack, M.P., George, R.C., 2009. Genetic characteristics of pneumococcal disease in elderly patients before introducing the pneumococcal conjugate vaccine. Epidemiol. Infect. 137, 1049–1056.

Pizza, M., Scarlato, V., Masignani, V., Giuliani, M.M., Arico, B., Comanducci, M., Jennings, G.T., Baldi, L., Bartolini, E., Capecchi, B., Galeotti, C.L., Luzzi, E., Manetti, R., Marchetti, E., Mora, M., Nuti, S., Ratti, G., Santini, L., Savino, S., Scarselli, M., Storni, E., Zuo, P., Broeker, M., Hundt, E., Knapp, B., Blair, E., Mason, T., Tettelin, H., Hood, D.W., Jeffries, A.C., Saunders, N.J., Granoff, D.M., Venter, J.C., Moxon, E.R., Grandi, G., Rappuoli, R., 2000. Identification of vaccine candidates against serogroup B *meningococcus* by whole-genome sequencing. Science 287, 1816–1820.

Plucinski, M.M., Starfield, R., Almeida, R.P., 2011. Inferring social network structure from bacterial sequence data. PLoS ONE 6, e22685.

Posada, D., Buckley, T.R., 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53, 793–808.

Posada, D., Crandall, K.A., 2001. Intraspecific gene genealogies: trees grafting into networks. Trends Ecol. Evol. 16, 37–45.

Posada, D., Crandall, K.A., 2002. The effect of recombination on the accuracy of phylogeny estimation. J. Mol. Evol. 54, 396–402.

Posada, D., Crandall, K.A., Holmes, E.C., 2002. Recombination in evolutionary genomics. Annu. Rev. Genet. 36, 75–97.

Pourcel, C., Andre-Mazeaud, F., Neubauer, H., Ramisse, F., Vergnaud, G., 2004. Tandem repeats analysis for the high resolution phylogenetic analysis of *Yersinia pestis*. BMC Microbiol. 4, 22.

Racloz, V.N., Luiz, S.J., 2010. The elusive meningococcal meningitis serogroup: a systematic review of serogroup B epidemiology. BMC Infect. Dis. 10, 175.

Rambaut, A., Drummond, A.J., 2009. Tracer: MCMC Trace Analysis Tool, 1.4.1 ed. Institute of Evolutionary Biology, Edinburgh. <http://tree.bio.ed.ac.uk/software/tracer/>.

Raymond, B., Wyres, K.L., Sheppard, S.K., Ellis, R.J., Bonsall, M.B., 2010. Environmental factors determining the epidemiology and population genetic structure of the *Bacillus cereus* group in the field. PLoS Pathog. 6, e1000905.

Ribeiro, F., Przybylski, D., Yin, S., Sharpe, T., Gnerre, S., Abouelleil, A., Berlin, A.M., Montmayeur, A., Shea, T.P., Walker, B.J., Young, S.K., Russ, C., MacCallum, I., Nusbaum, C., Jaffe, D.B., 2012. Finished bacterial genomes from shotgun sequence data. Genome Research doi. http://dx.doi.org/10.1101/gr.141515.112.

Robinson, D.A., Monk, A.B., Cooper, J.E., Feil, E.J., Enright, M.C., 2005. Evolutionary genetics of the accessory gene regulator (agr) locus in *Staphylococcus aureus*. J. Bacteriol. 187, 8312–8321.

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.

Rosenberg, M.S., 2009. Sequence Alignment. University of California Press, Berkeley, CA, p. 337.

Russell, J.A., Goldman-Huertas, B., Moreau, C.S., Baldo, L., Stahlhut, J.K., Werren, J.H., Pierce, N.E., 2009. Specialization and geographic isolation among *Wolbachia* symbionts from ants and lycaenid butterflies. Evolution 63, 624–640.

Sahin, O., Fitzgerald, C., Stroika, S., Zhao, S., Sippy, R.J., Kwan, P., Plummer, P.J., Han, J., Yaeger, M.J., Zhang, Q., 2012. Molecular evidence for zoonotic transmission of an emergent, highly pathogenic *Campylobacter jejuni* clone in the United States. J. Clin. Microbiol. 50, 680–687.

Sahl, J.W., Matalka, M.N., Rasko, D.A., 2012. Phylomark, a tool to identify conserved phylogenetic markers from whole-genome alignments. Appl. Environ. Microbiol. 78, 4884–4892.

Sakwinska, O., Giddey, M., Moreillon, M., Morisset, D., Waldvogel, A., Moreillon, P., 2011. *Staphylococcus aureus* host range and human-bovine host shift. Appl. Environ. Microbiol. 77, 5908–5915.

Salazar-Gonzalez, J.F., Bailes, E., Pham, K.T., Salazar, M.G., Guffey, M.B., Keele, B.F., Derdeyn, C.A., Farmer, P., Hunter, E., Allen, S., Manigart, O., Mulenga, J., Anderson, J.A., Swanstrom, R., Haynes, B.F., Athreya, G.S., Korber, B.T., Sharp, P.M., Shaw, G.M., Hahn, B.H., 2008. Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. J. Virol. 82, 3952–3970.

Schatz, M.C., Delcher, A.L., Salzberg, S.L., 2010. Assembly of large genomes using second-generation sequencing. Genome Res. 20, 1165–1173.

Schierup, M.H., Hein, J., 2000. Consequences of recombination on traditional phylogenetic analysis. Genetics 156, 879–891.

Schmidlin, M., Alt, M., Vogel, G., Voegeli, U., Brodmann, P., Bagutti, C., 2010. Contaminations of laboratory surfaces with *Staphylococcus aureus* are affected by the carrier status of laboratory staff. J. Appl. Microbiol. 109, 1284–1293.

Schouls, L.M., Van Der Ende, A., Damen, M., Van De Pol, I., 2006. Multiple-locus variable-number tandem repeat analysis of *Neisseria meningitidis* yields groupings similar to those obtained by multilocus sequence typing. J. Clin. Microbiol. 44, 1509–1518.

Schulte, P.A., Perera, F., 1993. Molecular Epidemiology: Principles and Practices. Academic Press, New York, NY.

Schultsz, C., Jansen, E., Keijzers, W., Rothkamp, A., Duim, B., Wagenaar, J.A., van der Ende, A., 2012. Differences in the population structure of invasive *Streptococcus* *suis* strains isolated from pigs and from humans in The Netherlands. PLoS ONE 7, e33854.

Schürch, A.C., van Soolingen, D., 2012. DNA fingerprinting of *Mycobacterium tuberculosis*: from phage typing to whole-genome sequencing. Infect. Genet. Evol. 12, 602–609.

Sheppard, S.K., Colles, F., Richardson, J., Cody, A.J., Elson, R., Lawson, A., Brick, G., Meldrum, R., Little, C.L., Owen, R.J., Maiden, M.C., McCarthy, N.D., 2010a. Host association of *Campylobacter* genotypes transcends geographic variation. Appl. Environ. Microbiol. 76, 5269–5277.

Sheppard, S.K., Dallas, J.F., Wilson, D.J., Strachan, N.J.C., McCarthy, N.D., Jolley, K.A., Colles, F.M., Rotariu, O., Ogden, I.D., Forbes, K.J., 2010b. Evolution of an agriculture-associated disease causing *Campylobacter coli* clade: evidence from national surveillance data in Scotland. PLoS ONE 5, e15708.

Simoes, R.R., Aires-de-Sousa, M., Conceicao, T., Antunes, F., da Costa, P.M., de Lencastre, H., 2010. High prevalence of EMRSA-15 in Portuguese public buses: a worrisome finding. PLoS ONE 6, e17630.

Singh, P., Foley, S.L., Nayak, R., Kwon, Y.M., 2012. Multilocus sequence typing of *Salmonella* strains by high-throughput sequencing of selectively amplified target genes. J. Microbiol. Methods 88, 127–133.

Smith, J.M., Smith, N.H., O'Rourke, M., Spratt, B.G., 1993. How clonal are bacteria? Proc. Natl. Acad. Sci. U.S.A. 90, 4384–4388.

Smith, M.A., Bertrand, C., Crosby, K., Eveleigh, E.S., Fernandez-Triana, J., Fisher, B.L., Gibbs, J., Hajibabaei, M., Hallwachs, W., Hind, K., Hrcek, J., Huang, D.W., Janda, M., Janzen, D.H., Li, Y., Miller, S.E., Packer, L., Quicke, D., Ratnasingham, S., Rodriguez, J., Rougerie, R., Shaw, M.R., Sheffield, C., Stahlhut, J.K., Steinke, D., Whitfield, J., Wood, M., Zhou, X., 2012. *Wolbachia* and DNA barcoding insects: patterns, potential, and problems. PLoS ONE 7, e36514.

Soge, O.O., Meschke, J.S., No, D.B., Roberts, M.C., 2009. Characterization of methicillin-resistant *Staphylococcus aureus* and methicillin-resistant coagulase-negative *Staphylococcus spp.* isolated from US West Coast public marine beaches. J. Antimicrob. Chemother. 64, 1148–1155.

Spratt, B.G., 1999. Multilocus sequence typing: molecular typing of bacterial pathogens in an era of rapid DNA sequencing and the internet. Curr. Opin. Microbiol. 2, 312–316.

Springman, A.C., Lacher, D.W., Wu, G., Milton, N., Whittam, T.S., Davies, H.D., Manning, S.D., 2009. Selection, recombination, and virulence gene diversity among group B streptococcal genotypes. J. Bacteriol. 191, 5419–5427.

Sproston, E.L., Ogden, I.D., MacRae, M., Dallas, J.F., Sheppard, S.K., Cody, A.J., Colles, F.M., Wilson, M.J., Forbes, K.J., Strachan, N.J., 2011. Temporal variation and host association in the *Campylobacter* population in a longitudinal ruminant farm study. Appl. Environ. Microbiol. 77, 6579–6586.

Sreevatsan, S., Pan, X., Stockbauer, K.E., Connell, N.D., Kreiswirth, B.N., Whittam, T.S., Musser, J.M., 1997. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. Proc. Natl. Acad. Sci. U.S.A. 94, 9869–9874.

Stabler, R.A., Dawson, L.F., Valiente, E., Cairns, M.D., Martin, M.J., Donahue, E.H., Riley, T.V., Songer, J.G., Kuijper, E.J., Dingle, K.E., Wren, B.W., 2012. Macro and micro diversity of *Clostridium difficile* isolates from diverse sources and geographical locations. PLoS ONE 7, e31559.

Stackebrandt, E., Frederiksen, W., Garrity, G.M., Grimont, P.A.D., Peter, K., Maiden, M.C.J., Nesme, X., Rossell, R., Swings, J., Tr, H.G., 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int. J. Syst. Evol. Microbiol. 52, 1043–1047.

Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.

Stefanelli, P., Fazio, C., Sofia, T., Neri, A., Mastrantonio, P., 2009. Serogroup C *meningococci* in Italy in the era of conjugate menC vaccination. BMC Infect. Dis. 9, 135.

Stensvold, C.R., Alfellani, M., Clark, C.G., 2012. Levels of genetic diversity vary dramatically between *Blastocystis* subtypes. Infect. Genet. Evol. 12, 263–273.

Stumpf, M.P.H., McVean, G.A.T., 2003. Estimating recombination rates from population-genetic data. Nat. Rev. Genet. 4, 959–968.

Tanigawa, K., Watanabe, K., 2011. Multilocus sequence typing reveals a novel subspeciation of *Lactobacillus delbrueckii*. Microbiology 157, 727–738.

Tay, C.Y., Mitchell, H., Dong, Q., Goh, K.L., Dawes, I.W., Lan, R., 2009. Population structure of *Helicobacter pylori* among ethnic groups in Malaysia: recent acquisition of the bacterium by the Malay population. BMC Microbiol. 9, 126.

Taylor, C.F., 2009. Mutation scanning using high-resolution melting. Biochem. Soc. Trans. 37, 433–437.

Tazi, L., Pérez-Losada, M., Gu, W., Yang, Y., Xue, L., Crandall, K.A., Viscidi, R.P., 2010. Population dynamics of *Neisseria gonorrhoeae* in Shanghai, China: a comparative study. BMC Infect. Dis. 10, 13.

Templeton, A.R., Crandall, K.A., Sing, C.F., 1992. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data III. Cladogram estimation. Genetics 132, 619–633.

Top, J., Schouls, L.M., Bonten, M.J.M., Willems, R.J.L., 2004. Multiple-locus variable-number tandem repeat analysis, a novel typing scheme to study the genetic relatedness and epidemiology of *Enterococcus faecium* isolates. J. Clin. Microbiol. 42, 4503–4511.

Torpdahl, M., Skov, M.N., Sandvang, D., Baggesen, D.L., 2005. Genotypic characterization of *Salmonella* by multilocus sequence typing, pulsed-field gel electrophoresis and amplified fragment length polymorphism. J. Microbiol. Methods 63, 173–184.

Trotter, C.L., Chandra, M., Cano, R., Larrauri, A., Ramsay, M.E., Brehony, C., Jolley, K.A., Maiden, M.C., Heuberger, S., Frosch, M., 2007. A surveillance network for meningococcal disease in Europe. FEMS Microbiol. Rev. 31, 27–36.

Unemo, M., Dillon, J.A.R., 2011. Review and international recommendation of methods for typing *Neisseria gonorrhoeae* isolates and their Implications for improved knowledge of gonococcal epidemiology, treatment, and biology. Clin. Microbiol. Rev. 24, 447–458.

Urwin, R., Maiden, M.C.J., 2003. Multi-locus sequence typing: a tool for global epidemiology. Trends Microbiol. 11, 479–487.

Urwin, R., Russell, J.E., Thompson, E.A., Holmes, E.C., Feavers, I.M., Maiden, M.C., 2004. Distribution of surface protein variants among hyperinvasive *meningococci*: implications for vaccine design. Infect. Immun. 72, 5955–5962.

Van Berkum, P., Elia, P., Song, Q., Eardly, B.D., 2012. Development and application of a multilocus sequence analysis method for the identification of genotypes within genus *Bradyrhizobium* and for establishing nodule occupancy of soybean (*Glycine max* L. Merr). Mol. Plant–Microbe Interact. 25, 321–330.

Vanderkooi, O.G., Church, D.L., MacDonald, J., Zucol, F., Kellner, J.D., 2011. Community-based outbreaks in vulnerable populations of invasive infections caused by *Streptococcus pneumoniae* serotypes 5 and 8 in Calgary, Canada. PLoS ONE 6, e28547.

Vanlaere, E., Lipuma, J.J., Baldwin, A., Henry, D., De Brandt, E., Mahenthiralingam, E., Speert, D., Dowson, C., Vandamme, P., 2008. *Burkholderia latens sp. nov*., *Burkholderia diffusa sp. nov*., *Burkholderia arboris sp. nov*., *Burkholderia seminalis sp. nov*. and *Burkholderia metallica sp. nov*., novel species within the *Burkholderia cepacia* complex. Int. J. Syst. Evol. Microbiol. 58, 1580–1590.

Vanlaere, E., Baldwin, A., Gevers, D., Henry, D., De Brandt, E., LiPuma, J.J., Mahenthiralingam, E., Speert, D.P., Dowson, C., Vandamme, P., 2009. Taxon K, a complex within the *Burkholderia cepacia* complex, comprises at least two novel species, *Burkholderia contaminans sp. nov.* and *Burkholderia lata sp. nov*. Int. J. Syst. Evol. Microbiol. 59, 102–111.

Vergnaud, G., Pourcel, C., 2006. Multiple Locus VNTR (variable number of tandem repeat) analysis. Molecular Identification, Systematics, and Population Structure of Prokaryotes. Springer-Verlag, Berlin, Germany, pp. 83–104.

Vogel, U., Szczepanowski, R., Claus, H., Junemann, S., Prior, K., Harmsen, D., 2012. Ion torrent personal genome machine sequencing for genomic typing of *Neisseria meningitidis* for rapid determination of multiple layers of typing information. J. Clin. Microbiol. 50, 1889–1894.

Vos, M., 2011. A species concept for bacteria based on adaptive divergence. Trends Microbiol. 19, 1–7.

Vos, M., Didelot, X., 2009. A comparison of homologous recombination rates in bacteria and archaea. ISME J. 3, 199–208.

Walker, A.S., Eyre, D.W., Wyllie, D.H., Dingle, K.E., Harding, R.M., O'Connor, L., Griffiths, D., Vaughan, A., Finney, J., Wilcox, M.H., Crook, D.W., Peto, T.E., 2012. Characterisation of *Clostridium difficile* hospital ward-based transmission using extensive epidemiological data and molecular typing. PLoS Med. 9, e1001172.

Walther, B., Hermes, J., Cuny, C., Wieler, L.H., Vincze, S., Abou Elnaga, Y., Stamm, I., Kopp, P.A., Kohn, B., Witte, W., Jansen, A., Conraths, F.J., Semmler, T., Eckmanns, T., Lubke-Becker, A., 2012. Sharing more than friendship–nasal colonization with coagulase-positive *Staphylococci* (CPS) and co-habitation aspects of dogs and their owners. PLoS ONE 7, e35197.

Wang, Y., Rannala, B., 2008. Bayesian inference of fine-scale recombination rates using population genomic data. Philos. Trans. R. Soc. Lond. B 363, 3921–3930.

Wang, Y., Rannala, B., 2009. Population genomic inference of recombination rates and hotspots. Proc. Natl. Acad. Sci. U.S.A. 106, 6215–6219.

Waples, R.S., Gaggiotti, O., 2006. Invited review: what is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Mol. Ecol. 15, 1419–1439.

Weinert, L.A., Welch, J.J., Suchard, M.A., Lemey, P., Rambaut, A., Fitzgerald, J.R., 2012. Molecular dating of human-to-bovid host jumps by *Staphylococcus aureus* reveals an association with the spread of domestication. Biol. Lett. 8, 829–832.

Wilson, D.J., McVean, G., 2006. Estimating diversifying selection and functional constraint in the presence of recombination. Genetics 172, 1411–1425.

Woolley, S.M., Posada, D., Crandall, K.A., 2008. A comparison of phylogenetic network methods using computer simulation. PLoS ONE 3, e1913.

Xie, W., Lewis, P.O., Fan, Y., Kuo, L., Chen, M.-H., 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst. Biol. 60, 150–160.

Yan, Y., Cui, Y., Han, H., Xiao, X., Wong, H.C., Tan, Y., Guo, Z., Liu, X., Yang, R., Zhou, D., 2011. Extended MLST-based population genetics and phylogeny of *Vibrio parahaemolyticus* with high levels of recombination. Int. J. Food Microbiol. 145, 106–112.

Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591.

Yang, Z., Wong, W.S., Nielsen, R., 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22, 1107–1118.

Yeo, M., Mauricio, I.L., Messenger, L.A., Lewis, M.D., Llewellyn, M.S., Acosta, N., Bhattacharyya, T., Diosque, P., Carrasco, H.J., Miles, M.A., 2011. Multilocus sequence typing (MLST) for lineage assignment and high resolution diversity studies in *Trypanosoma cruzi*. PLoS Negl. Trop. Dis. 5, e1049.

Zeigler, D.R., 2003. Gene sequences useful for predicting relatedness of whole genomes in bacteria. Int. J. Syst. Evol. Microbiol. 53, 1893–1900.

Zwickl, D.J., 2006. Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets Under the Maximum Likelihood Criterion, Department of Biological Sciences. The University of Texas at Austin, Austin, TX.