Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Research paper

Microbial sequence typing in the genomic era

Marcos Pérez-Losada[a,b,c,*], Miguel Arenas[d], Eduardo Castro-Nallar[e]

[a] Computational Biology Institute, Milken Institute School of Public Health, George Washington University, Ashburn, VA 20147, USA
[b] CIBIO-InBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, Campus Agrário de Vairão, Vairão 4485-661, Portugal
[c] Children's National Medical Center, Washington, DC 20010, USA
[d] Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo, Spain
[e] Universidad Andrés Bello, Center for Bioinformatics and Integrative Biology, Facultad de Ciencias Biológicas, Santiago 8370146, Chile

ARTICLE INFO

Keywords:
Bacteria
Epidemiology
MLST
Pathogen
Typing
WGS

ABSTRACT

Next-generation sequencing (NGS), also known as high-throughput sequencing, is changing the field of microbial genomics research. NGS allows for a more comprehensive analysis of the diversity, structure and composition of microbial genes and genomes compared to the traditional automated Sanger capillary sequencing at a lower cost. NGS strategies have expanded the versatility of standard and widely used typing approaches based on nucleotide variation in several hundred DNA sequences and a few gene fragments (MLST, MLVA, rMLST and cgMLST). NGS can now accommodate variation in thousands or millions of sequences from selected amplicons to full genomes (WGS, NGMLST and HiMLST). To extract signals from high-dimensional NGS data and make valid statistical inferences, novel analytic and statistical techniques are needed. In this review, we describe standard and new approaches for microbial sequence typing at gene and genome levels and guidelines for subsequent analysis, including methods and computational frameworks. We also present several applications of these approaches to some disciplines, namely genotyping, phylogenetics and molecular epidemiology.

1. Introduction

Microbial typing techniques are greatly enhancing our insights into microbial population epidemiology and microbial diversity and are widely used in diagnostics, genomics and pathogenesis related with microbiology research (Boers et al., 2012; Van Belkum, 2002). In fact, our ability to accurately distinguish among strains of infectious pathogens is crucial for efficient epidemiological and surveillance analysis, studying microbial population structure and dynamics and, ultimately, developing improved public health control strategies (Cooper and Feil, 2004). To achieve these goals, several molecular typing methods have been proposed that can identify isolates worldwide (global epidemiology) and/or in localized disease outbreaks (local epidemiology); see (Foley et al., 2009) for a review.

Since 1998, the established standard for molecular typing is Multilocus Sequence Typing (MLST) (Maiden et al., 1998), which has proven to be an effective method for characterizing bacterial isolates. However, next-generation sequencing (NGS) technologies are drastically changing the field of microbial genomics research (Forde and O'Toole, 2013). These new sequencing methods supply a range of applications (Glenn, 2011), allowing for a more comprehensive and in depth analysis of the diversity, structure and content of microbial genomes compared to traditional automated Sanger capillary sequencers at a substantially lower cost (Mardis, 2008; Metzker, 2010). The reader is referred to the following reviews for further information about NGS technologies (Goodwin et al., 2016; Mardis, 2013, 2017; Metzker, 2010). As a consequence, in March 2017, a total of 246,189 (complete and draft) bacterial and 6615 viral genomes have been deposited in the genomes online database *GOLD* (Mukherjee et al., 2017), of which 1041 correspond to metagenomic studies (i.e., the study of microbial communities directly in their natural environments). NGS platforms have proven to be effective tools for the re-sequencing and de novo sequencing reference microbial species and strains (pathogens and underrepresented taxa), but also assembling genomes of entire microbial communities of unculturable microbes (microbiotas) (Kyrpides et al., 2014; Sangwan et al., 2016; Sharon and Banfield, 2013).

Non-cultured organisms represent the vast majority of the total microbial diversity which exists in the world (Pace, 2009). Microbial genomic studies usually focus on microbial diversity and structure at the species (or strain) and community levels through targeted sequencing of gene amplicons (e.g., housekeeping genes, 16S/18S rRNA, ITS) or shotgun sequencing of (nearly) full genomes (Caporaso et al., 2012; Chun and Rainey, 2014; Kwong et al., 2015; MacCannell, 2013; Petrosino et al., 2009; Vincent et al., 2016).

[*] Corresponding author at: Computational Biology Institute, George Washington University, Innovation Hall, 45085 University Drive, Ashburn, VA 20147, USA.
E-mail address: mlosada@gwu.edu (M. Pérez-Losada).

http://dx.doi.org/10.1016/j.meegid.2017.09.022
Received 22 May 2017; Received in revised form 18 September 2017; Accepted 19 September 2017

NGS strategies have expanded the versatility of widely used typing approaches, such as MLST, to accommodate high-throughput data (e.g., NGMLST and HiMLST). The drawback is that this massive amount of data comes in the form of short reads with relatively high sequencing errors; hence one needs to invest heavily in computational analysis. Additionally, novel analytic and statistical techniques that can handle these data had to be developed. Over the last five years new typing approaches that take advantage of parallel amplicon and whole genome sequencing have been proposed.

In this review, we describe and compare classical (e.g., MLST) and recent (e.g., HiMLST) DNA typing approaches (Section 2) of microbial sequence typing. Then we present statistical methods and computational tools for the analysis of nucleotide, locus, and genome data generated via Sanger and NGS platforms (Section 3). In the last section (Section 4), we present several applications of these DNA-based approaches to the fields of genotyping, phylogenetics and molecular epidemiology. We also refer the reader to other reviews on MLST for complementary information (Boers et al., 2012; Cooper and Feil, 2004; Jolley et al., 2012; Jolley and Maiden, 2014; Larsen et al., 2012; Maiden, 2006; Pérez-Losada et al., 2017; Pérez-Losada et al., 2006; Pérez-Losada et al., 2013; Sullivan et al., 2005; Urwin and Maiden, 2003).

2. DNA-based typing approaches

2.1. Standard typing approaches: MLST and MLVA

Multilocus Sequence Typing (*MLST*) examines nucleotide variation in sequences of internal fragments of usually seven housekeeping genes, i.e., those encoding fundamental metabolic functions; although the number of genes may vary in dependence of the strains, species, and other particularities of the studied sample. For each gene, then the different sequences present within a species are assigned as distinct alleles and, for each isolate, the alleles at each of the seven loci define the allelic profile or sequence type (ST). Each isolate is therefore unambiguously characterized by a series of seven integers, which correspond to the alleles at the seven housekeeping loci.

MLST is widely used for molecular typing (Jolley and Maiden, 2014; Maiden, 2006; Maiden et al., 2013; Pérez-Losada et al., 2013). This was made possible by three advances in molecular microbiology (Maiden, 2006) involving: 1) bacterial evolution and population biology knowledge; 2) high-throughput nucleotide sequencing; and 3) genetic sequence databases. The bacterial population studies undertaken from the 1980s onwards were central to the development of MLST. Those studies showed that genetic exchange among bacteria was more common than previously thought, leading to a reassessment of the role of sexual processes in the structuring of bacterial populations. Using sequence data, it has been shown that recombination (mosaic genes) was not only frequent in genes under diversifying selection (e.g., antigen-encoding and antibiotic resistant genes), but also in genes under purifying selection (housekeeping genes) (see Maiden, 2006). This suggested that the clonal model (variation can only arise by mutation) was not universal and led to the proposal of new non-clonal or panmictic (variation is mainly generated by recombination) and partially clonal models of bacterial population structure (Feil and Enright, 2004; Spratt and Maiden, 1999). Consequently, typing methods needed to accommodate and distinguish among a broader spectrum of population structures, hence providing not only discriminatory power but also information about the clonal structure of the organism under study. Therefore, only molecular techniques that can contrast results across independent markers (such as MLST) would be adequate for bacterial typing and population genetic analyses.

The length of the nucleotide sequence amplified for each locus is generally in the range of 400–600 bp and is determined largely by the parameters of automated sequencing instruments available at the time. Most MLST nucleotide sequence data are generated by Sanger sequencing, however this technology is being replaced by high-throughput technologies such as pyrosequencing (Boers et al., 2012; Margulies et al., 2005), sequencing-by-synthesis (Illumina/Ion Torrent) and single-molecule sequencing (PacBio/Nanopore) (Chen et al., 2015; Pérez-Losada et al., 2013) for targeted-amplicon and whole-genome sequencing. These technologies can generate accurate read lengths of ~150 bp to 10 kb (Illumina and PacBio, respectively) and up to 25–50 million paired-end reads (Illumina MiniSeq/MiSeq platforms) per run. Moreover, the design of barcoded primers allows simultaneous and efficient sequencing of homologous products from hundreds of samples in the same run (Kozich et al., 2013; Rao et al., 2016; Taylor et al., 2016).

One of the goals of the MLST approach was the development of online platforms containing MLST databases to which public health officials and researchers could both have access and contribute and from which clinical, epidemiological and population studies could benefit (Maiden, 2006; Maiden et al., 1998; Urwin and Maiden, 2003). The first MLST online platforms were based on single databases implemented in the MLSTdB software (Chan et al., 2001), but as MLST schemes began to expand several limitations became apparent: redundant information (each record contained the ST designation and the allelic profile), isolate bias (single databases were dominated by specific studies), and access (all databases were stored at a single location). To overcome these limitations, a new network-based database software, MLSTdBNet (Jolley et al., 2004) was developed and implemented on the PubMLST site (http://pubmlst.org/). This site includes two databases: *i*) a profiles database with the sequences of each MLST allele for each locus linked to an allele number, and *ii*) an allelic profiles database with the corresponding ST designations. The profile database can then interact with other isolate databases. For each scheme on the PubMLST site there is a PubMLST isolate database that aims to include at least one isolate for each ST. MLST databases are hence different to other repository databases such as GenBank, not only in organization but also in active curation for accuracy. It is important to highlight that MLST databases do not embody the global diversity of an organism but the extent of its diversity at the time of access. Moreover, stored data are unstructured and do not necessarily represent natural populations either. As high-throughput sequencing becomes more affordable, PubMLST is increasingly including whole genome sequences, e.g., BIGSdb (Jolley and Maiden, 2010; Larsen et al., 2012).

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

The number of loci that should be evaluated to confidently assign a ST has been minimized to reduce the expense and time required for characterization, with most studies using 6–10 loci. If performed manually, evaluating even these many loci can be time consuming. However, fully automated systems, e.g., robotics (Jefferies et al., 2003; Sullivan et al., 2006) provide a high-throughput pipeline for data collection that can run large volumes of samples with increased reliability. Likewise, commercial solutions such as Ion Torrent AmpliSeq panels targeting MLST schemes (www.ampliseq.com) can reduce costs down to cents per marker. As sequencing technology progresses, we expect the cost of automation to decrease, thus data interpretation rather than data generation will be the likely limiting factor in our understanding of pathogen population dynamics.