REVIEW 10.1111/1469-0691.12217

Rapid bacterial genome sequencing: methods and applications in clinical microbiology

C. Bertelli[1,2] **and G. Greub**[1]

1) *Centre for Research on Intracellular Bacteria, Institute of Microbiology, University Hospital Center and University of Lausanne, Lausanne, Switzerland and* 2) *SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland*

Abstract

The recent advances in sequencing technologies have given all microbiology laboratories access to whole genome sequencing. Providing that tools for the automated analysis of sequence data and databases for associated meta-data are developed, whole genome sequencing will become a routine tool for large clinical microbiology laboratories. Indeed, the continuing reduction in sequencing costs and the shortening of the 'time to result' makes it an attractive strategy in both research and diagnostics. Here, we review how high-throughput sequencing is revolutionizing clinical microbiology and the promise that it still holds. We discuss major applications, which include: (i) identification of target DNA sequences and antigens to rapidly develop diagnostic tools; (ii) precise strain identification for epidemiological typing and pathogen monitoring during outbreaks; and (iii) investigation of strain properties, such as the presence of antibiotic resistance or virulence factors. In addition, recent developments in comparative metagenomics and single-cell sequencing offer the prospect of a better understanding of complex microbial communities at the global and individual levels, providing a new perspective for understanding host–pathogen interactions. Being a high-resolution tool, high-throughput sequencing will increasingly influence diagnostics, epidemiology, risk management, and patient care.

Keywords: Antibiotic resistance, bacterial genome, diagnostic, high-throughput sequencing, virulence factor

Clin Microbiol Infect

Corresponding author: G. Greub, Institute of Microbiology, Bungnon 48, 1011 Lausanne, Switzerland
E-mail: Gilbert.Greub@chuv.ch

Introduction

In recent years, there has been a major transformation in the way that clinicians and researchers extract genomic information from patient samples. The development of ultra-high-throughput sequencing (UHTS) technologies has been instrumental in advancing research in all scientific areas, but particularly in microbiology, where genomes are small. As shown by the impressive increase in genomic data output (Fig. 1), whole genome sequencing (WGS) has entered all research laboratories, and will soon become an integrated tool in diagnostic laboratories.

In clinical bacteriology, it is critical to rapidly characterize the pathogen present in a clinical sample, to improve patient care. Identification at the species level and antibiotic susceptibility testing are of major importance in guiding antibiotic treatment and the management of infectious diseases. Although matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry (MS) has been a revolution in clinical microbiology, and may have significant applications in typing, identification, and even toxin detection [1], there is still no high-throughput approach with which to fully and rapidly characterize any bacterial strain. Generally, such detailed characterizations involved multiple analyses, and were only performed by research laboratories specializing in a given pathogen. These detailed analyses take days to months, depending on the type of bacterium and the complexity of the question.

UHTS offers the possibilities of reducing the number of steps needed for full characterization of the pathogen, and optimizing the 'time to result' (Fig. 2). Bacterial whole genome shotgun sequence data can be obtained from a pure culture,

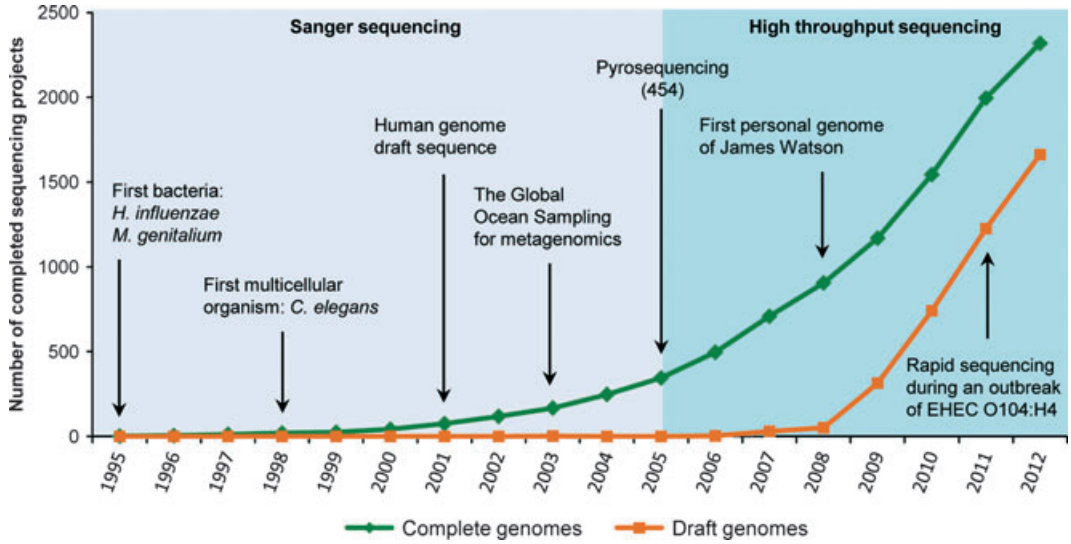

FIG. 1. Milestones in whole genome sequencing. The number of complete and draft genomes from the *Archaea* (n = 172), *Bacteria* (n = 3625) and *Eukarya* (n = 183) deposited in public databases from 1995 to 2012 is shown, as extracted from the GOLD database (www.genomesonline.org/) in December 2012. The advent of high-throughput sequencing led to a rapid increase in the number of complete genome sequences (green), but the short read length, in particular, triggered the final publication of draft genome sequences (orange). EHEC, enterohaemorrhagic *Escherichia coli*.

directly from a clinical sample, or even from a single bacterium present in a given sample [2,3]. Genome finishing—the most costly and time-consuming step—is often not necessary, and the release of unfinished genomes has become a major trend in the area [4,5]. Unfinished genomes, also called raw, draft or dirty genomes, provide enough data for extraction of the required information (Fig. 3), such as the presence of toxins [6] or genes or mutations coding for antibiotic resistance [7]. Unfinished genomes can also be directly used to develop new diagnostic tests such as ELISA [4] and PCR [8].

WGS has become rapid and cheap enough to replace some older techniques previously used to characterize a pathogen at the genomic level. We review here the main UHTS techniques, and discuss their main applications in clinical and diagnostic laboratories. Finally, we highlight substantial challenges that remain in the development of innovative pipelines for genome analysis and data storage, to gather information in an effective, accurate and harmonized way.

Sequencing Technologies—in Short

In 2005, new high-throughput sequencing technologies appeared on the market, and were referred to as 'next-generation sequencing' technologies, as they replaced Sanger's dideoxy chain termination sequencing method. Their development was quick and remarkable, and they rapidly turned out to be essential tools for microbial genomics. Next-generation sequencing technologies have been the subject of excellent reviews [9,10], and we will only highlight their main advantages and limitations with regard to their use in clinical microbiology (Table 1).

The 454 Genome Sequencer, the first to be commercialized, rapidly established itself as a standard for *de novo* sequencing and metagenomics, thanks to its long reads (up to 700 bp) [11]. Shortly thereafter, Solexa sequencing by synthesis became the most widely used system among the research community [12]. It has major applications in resequencing and RNA sequencing, thanks to its high throughput, allowing a lower cost per base, although with shorter reads (36–150 bp). The SOLiD sequencing system [13], based on two-base sequencing by ligation, is insensible to homopolymer errors, and is principally used for resequencing, transcriptomics, or epigenomics. Arriving later in the field, Heliscope [14] remained marginally used by the community. Both 454 and Solexa offer the possibility of obtaining paired-read information, which is of great help for assembly. In addition, samples can easily be tagged with short (6–8 bp) barcoding sequences and pooled into a single run. Finally, thanks to its higher throughput, Illumina enables the multiplexing and sequencing of nearly 100 bacterial samples at a time, making it a cost-effective platform for sequencing large collections of bacteria.

More recently, a third generation of technologies arrived on the market. Life Science launched the Ion Torrent PGM and the Ion Proton Sequencer, which are based on the sensing of proton release during base incorporation [15]. The major