International Journal of Infectious Diseases 56 (2017) 130–135

Contents lists available at ScienceDirect

International Journal of Infectious Diseases

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

Review

Removing the bottleneck in whole genome sequencing of *Mycobacterium tuberculosis* for rapid drug resistance analysis: a call to action

Ruth McNerney [a,*], Taane G. Clark [b], Susana Campino [b], Camilla Rodrigues [c], David Dolinger [d], Liezel Smith [a], Andrea M. Cabibbe [e], Keertan Dheda [a], Marco Schito [f]

[a] Lung Infection and Immunity Unit, Department of Medicine, Division of Pulmonology and UCT Lung Institute, University of Cape Town, Old Main Building, Groote Schuur Hospital, Observatory, Cape Town, South Africa
[b] Faculty of Infectious and Tropical Diseases and Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK
[c] Department of Microbiology, P.D. Hinduja National Hospital and Medical Research Centre, Mumbai, India
[d] FIND, Campus Biotech, Geneva, Switzerland
[e] Emerging Bacterial Pathogens Unit, Division of Immunology, Transplantation and Infectious Diseases, IRCCS San Raffaele Scientific Institute, Milan, Italy
[f] Critical Path Institute, Tucson, Arizona, USA

ARTICLE INFO

Article history:
Received 4 October 2016
Received in revised form 30 November 2016
Accepted 30 November 2016

Corresponding Editor: Eskild Petersen, Aarhus, Denmark

Keywords:
Tuberculosis
Next-generation sequencing
Drug resistance
Clinical isolate
Sputum

SUMMARY

Whole genome sequencing (WGS) can provide a comprehensive analysis of *Mycobacterium tuberculosis* mutations that cause resistance to anti-tuberculosis drugs. With the deployment of bench-top sequencers and rapid analytical software, WGS is poised to become a useful tool to guide treatment. However, direct sequencing from clinical specimens to provide a full drug resistance profile remains a serious challenge. This article reviews current practices for extracting *M. tuberculosis* DNA and possible solutions for sampling sputum. Techniques under consideration include enzymatic digestion, physical disruption, chemical degradation, detergent solubilization, solvent extraction, ligand-coated magnetic beads, silica columns, and oligonucleotide pull-down baits. Selective amplification of genomic bacterial DNA in sputum prior to WGS may provide a solution, and differential lysis to reduce the levels of contaminating human DNA is also being explored. To remove this bottleneck and accelerate access to WGS for patients with suspected drug-resistant tuberculosis, it is suggested that a coordinated and collaborative approach be taken to more rapidly optimize, compare, and validate methodologies for sequencing from patient samples.

1. Introduction

The publication of the first *Mycobacterium tuberculosis* genome sequence in 1998 heralded a new era in tuberculosis (TB) research.[1] The bacterium was found to have a relatively small circular genome of approximately 4.5 million base pairs and was estimated to have around 4000 potential genes. Sequencing of the *M. tuberculosis* H37Rv reference strain was soon followed by other strains, and expectations were high that mechanisms of pathogenesis and virulence were about to be revealed that would enable the development of novel drugs and improved vaccines to assist in TB control.[2,3] However, the anticipated advances have been slow to materialize and such work continues, accompanied by the realization of the sophistication of *M. tuberculosis* as a highly successful human pathogen.

The early genomes were deciphered using various cloning and shotgun sequencing approaches, followed by assembly and annotation with an assortment of analytical tools – a painstaking process with each genome taking years to complete. Advances in technology over the past 15 years have greatly reduced the complexity, cost, and time of sequencing. Thousands of *M. tuberculosis* strains have since been sequenced, contributing to studies on evolution, transmission, and drug resistance.[4–10] Several Web-based tools have been developed to assist TB sequence analysis, and software is now freely available for rapid genotypic analysis and the identification of drug resistance-associated mutations.[11–15]

* Corresponding author.
E-mail address: Ruth.Mcnerney@uct.ac.za (R. McNerney).

http://dx.doi.org/10.1016/j.ijid.2016.11.422

Next-generation sequencing (NGS) is already used in clinical practice for characterizing cancers and hereditary diseases. It may also be used to assist TB case management by detecting resistance to anti-TB drugs.[16–18] However, the quality of the DNA template is critical for successful whole genome sequencing (WGS), and *M. tuberculosis* presents considerable technical challenges in this respect. Firstly, the scarcity of bacilli in clinical samples limits the availability of *M. tuberculosis* genomic DNA (gDNA). Secondly, the remarkable hardiness of the lipid-rich *M. tuberculosis* cell wall makes disruption of the *M. tuberculosis* bacterium difficult and can affect the quality and the yield of gDNA. In addition, the *M. tuberculosis* genome itself is unusually robust, with a guanine/cytosine average content of 65% across the genome, and with some regions exceeding 80%.[1] Thus careful consideration must be given to the choice of sample, bacterial lysis, DNA extraction methodology, library preparation, and sequencing platform if *M. tuberculosis* sequencing is to enter routine clinical practice.[19]

This article explores sample collection and processing as crucial factors for WGS and NGS analysis of patient-derived *M. tuberculosis*. Although some questions remain unanswered, advice is offered on current best practices and pitfalls to avoid.

2. Next-generation sequencing platforms

NGS technologies analyse whole genomes without recourse to cloning, as was previously required for Sanger sequencing. High throughput platforms have been developed that can analyse millions of DNA fragments in parallel, and sequencing of bacterial genomes that previously took years to complete can now be achieved in hours. Sophisticated labelling systems allow multiplexing, where multiple samples are combined within a single run, greatly reducing costs. The early NGS platforms required specialized laboratories and highly trained personnel, but newer bench-top instruments suitable for clinical laboratories are now available. In addition to WGS, targeted sequencing, where multiple targets are amplified and sequenced in parallel, offers a rapid detection of resistance, albeit on a limited scale compared to whole genome approaches.[17,20]

Several NGS systems are commercially available, providing a range of platforms to choose from, some of which have gone through regulatory registration and clearance. Two NGS systems have been used to detect mutations associated with *M. tuberculosis* drug resistance.[21] Illumina sequencing (Illumina, San Diego, CA, USA) is based on reversible dye-terminators.[22] First a library is prepared whereby pure DNA is chopped into smaller fragments that are modified prior to amplification on a specialized chip holding hundreds of thousands of oligonucleotides. New fragments are built one nucleotide at a time, with fluorescent tags indicating which nucleotide has been incorporated. Sequential rounds of nucleotide additions build a new strand of DNA, with thousands of positions throughout the genome being sequenced at the same time in a process called massive parallel sequencing. The result is millions of short DNA fragments replicating the entire genome, with multiple fragments covering each stretch of gDNA. In most cases fragments are aligned and assembled by computation against a predetermined reference genome, thus allowing for the identification of polymorphisms. The depth of coverage (number of fragments that represents a specific nucleotide) is an indication of sequencing quality for each nucleotide sequenced, because if hundreds of fragments give the same signal for a specific position there is high confidence in the specific nucleotide call. Depth of coverage varies across genomes and is adversely affected by regions of high GC content and repetitive elements. Typically, when calling drug resistance mutations, coverage of at least 10-fold is required. However, other studies suggest at least 30-fold coverage is needed, else the result is considered of low confidence.[19]

Ion Torrent or Ion semiconductor sequencing (Ion Torrent Inc., USA, marketed by Thermo Fisher Scientific) uses different chemistry to monitor nucleotides incorporated during the creation by polymerization reaction of the new strand of DNA.[23] Sequencing is performed on semiconductor chips that detect changes in pH caused by the release of H^+ ions during the polymerization of DNA and incorporation of deoxyribonucleoside triphosphate (dNTP). dNTPs are added sequentially, with washing between each step to remove unbound nucleotide molecules. As with Illumina sequencing, assembly by computation and comparison with reference genomes provides a readout of the whole genome and identifies polymorphisms. Labelling dyes and optical detection are not employed in this platform, a factor that the manufacturer claims increases the speed of sequencing and reduces running costs. Ion Torrent sequencing has been used successfully to detect mutations associated with *M. tuberculosis* drug resistance.[16,21]

PacBio sequencing, or single molecule real-time sequencing (Pacific Biosciences, USA), is primarily being utilized as a research tool.[24] Rather than short fragments, this technology generates long strands of DNA and it is possible to assemble whole bacterial genomes from a single reaction. Assembly problems for repetitive regions encountered with short-read sequencing methods are overcome, for example the *M. tuberculosis* PE and PPE gene families, which are often excluded from Illumina sequencing data analysis. The technology may also be used to investigate DNA methylation, DNA damage, and other epigenetic information. PacBio sequencing is rapid, but has a limited capacity to multiplex, which translates to a moderate throughput. A drawback when considering clinical applications is the requirement for large amounts of pure high molecular mass gDNA, which for *M. tuberculosis* necessitates lengthy culture and extraction processes.

The MinION sequencing platform (Oxford Nanopore Technologies, UK) is a rapid technology with a portable sequencing instrument that also offers relatively long read lengths.[25] It was recently shown to detect antibiotic resistance genes of *Escherichia coli* in DNA purified from urine following processing to remove contaminating human DNA, with results available in approximately 4 h.[26] However, successful sequencing of the *M. tuberculosis* genome with this platform has not yet been reported.

The choice of sequencing strategy is influenced by the reasoning for sequencing. Identifying novel polymorphisms of biological significance ideally requires a depth of coverage of at least 100-fold, and for discovering rare polymorphisms up to a 1000-fold coverage may be preferred. Similarly, to investigate transmission chains and differentiate between closely related bacteria with confidence requires high coverage. Lower coverage (20- to 30-fold) may be sufficient to observe specific well-defined single nucleotide polymorphisms (SNPs). Thus, while high quality DNA with few strand breaks may be crucial for some research applications, DNA of lower quality may be tolerated when looking for known SNPs associated with drug resistance. For sequencing to be optimally effective for managing patients with drug-resistant TB, results should be available in days rather than weeks and ideally sequencing should be performed directly from primary clinical specimens. It should be noted, however, that when dealing with patient samples, mixed infections (more than one *M. tuberculosis* strain) and heteroresistance (more than one polymorphism at the same point) may require special consideration, and higher depth of reads are desirable. There is as yet no consensus on the depth of coverage required for NGS drug resistance screening for *M. tuberculosis*.

3. Sputum collection and processing

The sample of choice for the diagnosis of pulmonary TB is expectorated sputum, which is subjected to analysis by smear