




# Whole genome sequencing of *Mycobacterium tuberculosis*

Andrea M. Cabibbe <sup>1</sup>, Timothy M. Walker<sup>2</sup>, Stefan Niemann<sup>3</sup> and Daniela M. Cirillo<sup>1</sup>

**Affiliations:** <sup>1</sup>Emerging Bacterial Pathogens Unit, IRCCS San Raffaele Scientific Institute, Milan, Italy. <sup>2</sup>Nuffield Dept of Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK. <sup>3</sup>Molecular and Experimental Mycobacteriology, Research Center Borstel, Borstel, Germany.

**Correspondence:** Daniela M. Cirillo, Emerging Bacterial Pathogens Unit, Division of Immunology, Transplantation and Infectious Diseases, IRCCS Ospedale San Raffaele, Via Olgettina 58, 20132 Milan, Italy. E-mail: cirillo.daniela@hsr.it

 @ERSpublications

Next-generation sequencing approaches promise to become the future standard for drug-susceptibility testing and epidemiological investigation in tuberculosis and for other high-priority bacterial pathogens <http://ow.ly/kXsV30ILP78>

**Cite this article as:** Cabibbe AM, Walker TM, Niemann S, *et al.* Whole genome sequencing of *Mycobacterium tuberculosis*. *Eur Respir J* 2018; 52: 1801163 [<https://doi.org/10.1183/13993003.01163-2018>].

Next-generation sequencing (NGS) technologies using massively parallel processing to interrogate pathogen genomes in days are revolutionising clinical microbiology practice [1]. Whole genome sequencing (WGS) offers unprecedented resolution for genotyping, outbreak investigation and determination of known sequence variants involved in antimicrobial resistance, which deep sequencing of selected genomic regions (targeted NGS) can further illuminate. WGS-based approaches have been proposed for surveillance of bacterial pathogens included in the “priority list” by the World Health Organization (WHO) [2]. At present, proof-of-principle and validation studies have been conducted for WGS from culture samples of *Escherichia coli*, *Klebsiella pneumoniae*, *Staphylococcus aureus*, *Streptococcus pneumoniae*, *Pseudomonas aeruginosa*, *Salmonella* spp., *Acinetobacter* spp., *Neisseria gonorrhoeae* and *Clostridium difficile* included in this list, in addition to the globally established priority group *Mycobacterium* spp. (including *Mycobacterium tuberculosis*) [3].

## Sequencing technologies

For years, clinical microbiology laboratories have relied on the use of conventional phenotypic and genotypic methods for species identification, detection of antibiotic resistance and studying transmission of bacteria responsible for hospital- or community-acquired infections. These are useful methods but they have long turnaround times, require specific infrastructure and have low discriminatory power to distinguish highly genetically related strains [4]. Moreover, technological constraints mean that only a few genomic loci can be simultaneously investigated [4]. Sanger DNA sequencing, developed in the 1970s, enabled the first gene and genome sequences but remained limited in its application due to the complexity and high sequencing costs when applied to extended genomic regions [5]. The sequencing output increased tremendously with the advent of pyrosequencing and then of NGS methods in the mid-2000s when costs plummeted [6]. By the end of the sequencing reaction, the NGS platforms generate the sequences of unique DNA fragments, known as “reads”, which can either be short (50–400 base pairs in length) or long (1–100 kilobases). A single NGS run can produce billions of sequence reads [5]. The workflow is similar for all the technologies: extraction of high molecular weight DNA; preparation of

---

Received: June 20 2018 | Accepted after revision: Sept 02 2018

Copyright ©ERS 2018

libraries (*i.e.* collection of DNA fragments) through DNA fragmentation (enzymatic or mechanical), barcoding and PCR amplification; clustering (clonal amplification of single DNA fragments); and automated single- or paired-end (*i.e.* the sequencer reads both ends of a DNA fragment) sequencing [5]. Among the major players in the field, Illumina Inc. (San Diego, CA, USA) uses sequencing by synthesis (bridge PCR) chemistry and offers a wide range of “benchtop” (MiniSeq and MiSeq) or high-throughput (NextSeq 500, HiSeq 2500 and Novaseq) solutions allowing higher output at lower costs for generation of high-quality short reads. Thermo Fisher Scientific Inc. (Waltham, MA, USA) “benchtop” (PGM and S5) and high-throughput (Ion Proton) platforms are based on sequencing by synthesis-semiconductor (emulsion PCR) chemistry, generating lower throughput but longer reads and in a shorter time, compared to Illumina, thus being well suited for targeted solutions, but with higher error rates in homopolymers [5]. The so-called third-generation sequencing platforms (RSII and Sequel (Pacific Biosciences of California Inc., Menlo Park, CA, USA) and MinION (Oxford Nanopore Technologies Limited, Oxford, UK)) take advantage of single-molecule real-time chemistry to generate long reads. These lend themselves to *de novo* assembly, which does not require a reference genome, and the sequencing of longer repetitive regions that short reads cannot span. Throughput and quality are, however, lower compared to instruments generating shorter reads [4, 5]. A detailed description of the main features and costs of different NGS technologies available for use in the microbiology field is reported in recent reviews [3–5, 7, 8]. The wide range of instruments marketed by NGS companies allows the user to adopt the most effective technology in the microbiology laboratory according to the purposes (routine, epidemiological and drug resistance surveillance, research), cost considerations (staff, capital investment, infrastructure, software/hardware), workload (facility, multi-disease platform, number of cases) and local availability.

### Role of sequencing in tuberculosis

Drug-resistant tuberculosis (TB) is considered a main cause of global morbidity and mortality related to antimicrobial resistance [9]. Multidrug-resistant (MDR) and extensively drug-resistant (XDR) forms are hampering TB control and clinical management, with treatment success of 54% and 30% for MDR- and XDR-TB, respectively (compared to 83% for susceptible forms) [9]. Phenotypic testing is the standard for drug-susceptibility testing (DST) despite its challenges, which are felt most acutely in resource-limited settings. These challenges include the slow growth rate of *M. tuberculosis*, the high-level biosafety infrastructure required, the poor reproducibility and some uncertainties around the proposed critical concentrations for some drugs [10]. As *M. tuberculosis* resistance emerges from changes in the genome, such as single nucleotide polymorphisms (SNPs), insertions and deletions in genes coding for drug targets or involved in drug metabolic pathways, or from efflux pump upregulation, methods for detecting genome variations represent a breakthrough for the rapid, simple and standardised management of drug-resistant TB cases [10]. There are a number of genotypic tests recommended by WHO for diagnosis of MDR- and XDR-TB, including cartridge-based nucleic acid amplification tests and line probe assays that can be implemented in peripheral TB laboratories [11]. However, these tools target only the “hot-spot” regions of a few genes to detect resistance to a restricted number of drugs, and do not always report the exact nucleotide change upon which a prediction of phenotypic resistance is based.

All-in-one solutions are needed to guide individualised clinical decisions for the most complicated resistant cases, at least at a reference laboratory level. NGS is the leading candidate technology in this regard [11]. The absence of integrative vectors and low mutation rate make the *M. tuberculosis* genome well suited for sequencing. The one technical challenge is the presence of repetitive and hard-to-sequence regions with high GC content. These require sufficient genome-wide sequencing depth to sequence more accurately, which has cost implications [10].

Multiple studies have used WGS to investigate TB resistance prediction, transmission dynamics and the population structure of *M. tuberculosis* complex [10, 12, 13]. In some settings, genome sequencing is already being used for the surveillance of disease transmission and the emergence of drug resistance in a population, as well as for resistance prediction for first- and second-line drugs [14–17]. Further work is being done on the discovery of targets involved in phenotypic resistance and the large-scale validation of resistance-conferring mutations [18]; on taxonomy [19]; on characterising hetero-resistance and mixed infections; and on virulence and pathogenesis [20]. Feasibility studies to evaluate the introduction of WGS into diagnostic algorithms of routine microbiology laboratories have been conducted in low-burden settings [21–23]. In such contexts, the rapid identification of drug-resistant TB and transmission events could help achieve TB elimination [24, 25]. Such feasibility studies have largely used Illumina technology, demonstrating how the integration of WGS from culture into the routine diagnostic workflow achieves accurate and timely reports, available several days earlier than phenotypic DST, at no greater costs.

The ongoing need for culturing poses a challenge to the full implementation of NGS as an effective alternative to conventional methods (*e.g.* Xpert MTB/RIF; Cepheid Inc., Sunnyvale, CA, USA), particularly

in resource-limited settings. Efforts are thus being made to develop protocols for WGS directly from clinical specimens. Such procedures include differential lysis steps, TB enrichment and automated DNA purification [26, 27]. These methods remain expensive at present and are challenged by the low starting material of *M. tuberculosis* and contamination with other genetic material (human and oral flora). Targeted approaches taking advantage of the selective amplification of phylogenetic and drug-resistance-related regions may represent a suitable alternative for direct sequencing (figure 1) [27].

A population-based surveillance study has been conducted in seven highly endemic countries with support of national and supranational reference laboratories, demonstrating that genetic sequencing targeting well-validated mutations [18] represents a powerful tool to determine/monitor resistance trends to first- and second-line drugs as a replacement for phenotypic tests that perform sub-optimally in such resource-limited settings [17]. Similarly, WGS is now proposed as the standard to detect transmission chains in real time and guide public health interventions at the highest resolution compared to traditional molecular methods, through either SNP-based (higher resolution) or core genome multilocus sequence typing (slightly less resolution) approaches [16].

### Sequencing implementation

Implementing NGS in TB clinical laboratories is complex and involves strategic planning, procurement, budgeting, sample referral systems, standard operating procedures, quality assurance, data management (storage, analysis, interpretation and reporting), and human resource (staff and training) considerations.

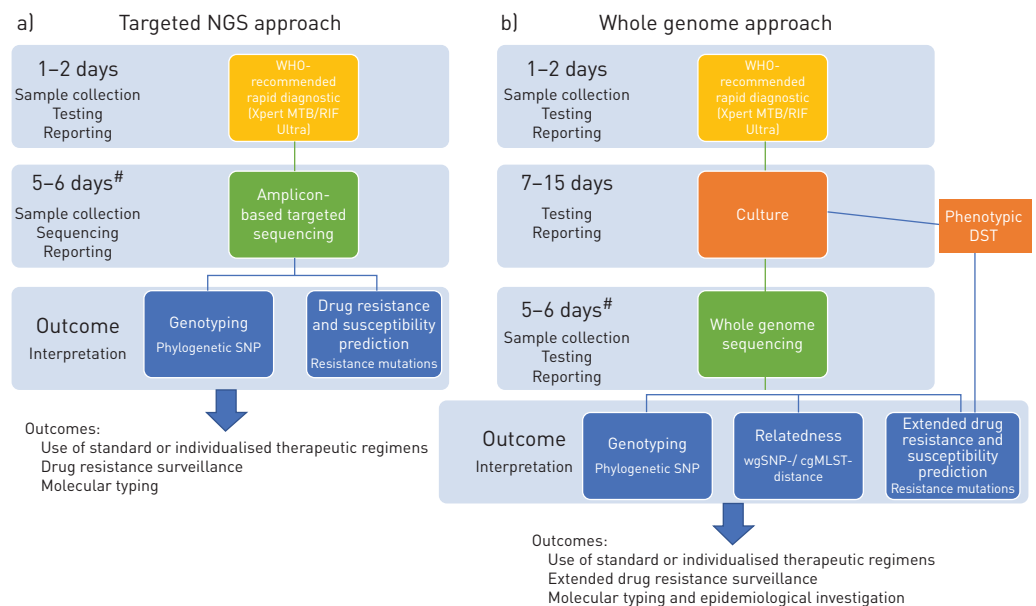


FIGURE 1 Whole genome sequencing of *Mycobacterium tuberculosis*. Two possible next-generation sequencing (NGS)-based scenarios and turnaround times to guide treatment decisions are presented. Modified Global Laboratory Initiative model tuberculosis (TB) diagnostic algorithms (revised June 2018 [28]) are outlined here and complemented with the NGS approach. a) Targeted NGS from clinical specimens (no culture required). b) Whole genome sequencing from cultured samples. Xpert MTB/RIF Ultra (Cepheid Inc., Sunnyvale, CA, USA) is used in this illustrative flow as the initial diagnostic test for persons being evaluated for TB. 1) If *M. tuberculosis* is detected but rifampicin resistance is not detected, treat with first-line regimen and refer sample for: a) targeted NGS, which may promptly reveal additional resistance and guide appropriate treatment [e.g. for rifampicin-susceptible, isoniazid-resistant TB]; or b) culture (and phenotypic drug-susceptibility testing (DST)), then as soon as the culture turns positive, perform whole genome sequencing, which may promptly reveal additional resistance and guide appropriate treatment [e.g. for rifampicin-susceptible, isoniazid-resistant TB], providing extended DST information. 2) If *M. tuberculosis* is detected and rifampicin resistance is detected, initiate treatment with second-line regimen or shorter multidrug-resistant TB (MDR-TB) regimen, as appropriate, and refer sample for: a) targeted NGS, which may promptly reveal additional resistance and guide standardised or individualised MDR/rifampicin-resistant (RR)-TB regimens; or b) culture (and phenotypic DST), then as soon as the culture turns positive, perform whole genome sequencing, which may promptly reveal additional resistance and guide standardised or individualised MDR/RR-TB regimens, providing extended DST information (including all repurposed and new drugs). The whole genome sequencing approach also enables high-resolution outbreak analysis to guide public health interventions. WHO: World Health Organization; SNP: single nucleotide polymorphism; wgSNP: whole genome SNP; cgMLST: core genome multilocus sequence typing. #: subjected to batching according to the NGS platform throughput.

Scale-up of sequencing laboratories requires adequate infrastructure: areas for sample preparation (DNA extraction from clinical isolates/specimens considering related biosafety); a molecular biology environment (pre- and post-PCR, and space for NGS instrument); power supply; favourable environment conditions (e.g. controlled temperature, humidity, vibration); network and internet connections; and computing capacity. Equipment and reagents required are platform specific: it is therefore essential to ensure availability of local distributors and prompt technical support. Several solutions are available for extraction and purification of genomic DNA, including commercially available (para)magnetic- and column-based systems, chemical procedures (e.g. CTAB (cetyltrimethyl ammonium bromide)/NaCl protocol), as well as devices for assessment of quantity/quality of the extracted DNA samples (fluorometer, spectrophotometer) [8]. Library preparation and sequencing reactions are performed according to the instructions provided by NGS manufacturers. Given the large amount of data generated by NGS, staff with a bioinformatics background are required to handle output data for storage and to run analytic pipelines. Large-scale validation of bioinformatic outputs is still needed and international consortia are putting huge effort into the standardisation of post-sequencing analysis processes for the reliable interpretation of drug-resistant TB profiles and epidemiological links. User-friendly solutions have been developed to avoid the need for specific bioinformatics skills that are both freely [29] or commercially available. As sequencing data require privacy-secured and long-lasting backup for next use, server- or cloud-based solutions are available for users. Staff require adequate and continuous training and mentoring on wet procedures and on post-sequencing processing.

### Next steps

To fully implement NGS into routine workflows, methods need to enter validation and certification programmes. Performance, accuracy and reproducibility, quality control steps, quality thresholds (e.g. on the depth/breadth of genome coverage), use of standards and development of standard operating procedures, impact on turnaround times and clinical management must all be assessed or evaluated [1, 7]. Microbiology laboratories introducing these technologies will undergo external proficiency testing programmes that are already implemented in TB for molecular testing. Simple but comprehensive clinical reports are crucial to help clinicians arrive at the best decisions in the management of TB cases. Although there is a huge amount of data generated by NGS, our knowledge as to all its meaning remains incomplete, and development of how best to report on this data in the meantime remains ongoing [30]. A report should at least give information on sequencing quality and identification of mutations to infer genotyping and drug resistance profiles, and provide details on the exact nucleotide changes and standardised prediction of resistance levels (ideally based on a literature review of minimum inhibitory concentration data).

WGS and targeted NGS approaches promise to become the future standard for DST and epidemiological investigation in TB, and for other high-priority bacterial pathogens [3, 31]. Additional work is needed to address the feasibility of WGS from clinical specimens, to standardise and automate the laboratory procedures and post-sequencing analyses, and to implement the NGS platforms in low-resource, high-burden settings.

Conflict of interest: A.M. Cabibbe has nothing to disclose. T.M. Walker has nothing to disclose. S. Niemann reports grants from the German Center for Infection Research and Leibniz Science Campus Evolung, during the conduct of the study. D.M. Cirillo has nothing to disclose.

### References

- 1 Rossen JWA, Friedrich AW, Moran-Gilad J, *et al.* Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. *Clin Microbiol Infect* 2018; 24: 355–360.
- 2 World Health Organization. Global Priority List of Antibiotic-resistant Bacteria to Guide Research, Discovery, and Development of New Antibiotics. Geneva, World Health Organization, 2017. Available from: [www.who.int/medicines/publications/global-priority-list-antibiotic-resistant-bacteria/en/](http://www.who.int/medicines/publications/global-priority-list-antibiotic-resistant-bacteria/en/)
- 3 Ellington MJ, Ekelund O, Aarestrup FM, *et al.* The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. *Clin Microbiol Infect* 2017; 23: 2–22.
- 4 Quainoo S, Coolen JPM, van Hijum SAFT, *et al.* Whole-genome sequencing of bacterial pathogens: the future of nosocomial outbreak analysis. *Clin Microbiol Rev* 2017; 30: 1015–1063.
- 5 Besser J, Carleton HA, Gerner-Smidt P, *et al.* Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clin Microbiol Infect* 2018; 24: 335–341.
- 6 Goldberg B, Sichtig H, Geyer C, *et al.* Making the leap from research laboratory to clinic: challenges and opportunities for next-generation sequencing in infectious disease diagnostics. *MBio* 2015; 6: e01888-15.
- 7 Kozyreva VK, Truong CL, Greninger AL, *et al.* Validation and implementation of clinical laboratory improvements act-compliant whole-genome sequencing in the public health microbiology laboratory. *J Clin Microbiol* 2017; 55: 2502–2520.

- 8 Association of Public Health Laboratories. Next Generation Sequencing Implementation Guide. Silver Spring, Association of Public Health Laboratories, 2016. [www.aphl.org/aboutAPHL/publications/Documents/ID-NGS-Implementation-Guide102016.pdf](http://www.aphl.org/aboutAPHL/publications/Documents/ID-NGS-Implementation-Guide102016.pdf)
- 9 World Health Organization. Global Tuberculosis Report 2017. Geneva, World Health Organization, 2017. Available from: [www.who.int/tb/publications/global\\_report/en/](http://www.who.int/tb/publications/global_report/en/)
- 10 McNerney R, Zignol M, Clark TG. Use of whole genome sequencing in surveillance of drug resistant tuberculosis. *Expert Rev Anti Infect Ther* 2018; 16: 433–442.
- 11 Cabibbe AM, Sotgiu G, Izco S, *et al.* Genotypic and phenotypic *M. tuberculosis* resistance: guiding clinicians to prescribe the correct regimens. *Eur Respir J* 2017; 50: 1702292.
- 12 Walker TM, Merker M, Kohl TA, *et al.* Whole genome sequencing for M/XDR tuberculosis surveillance and for resistance testing. *Clin Microbiol Infect* 2017; 23: 161–166.
- 13 Hatherell HA, Colijn C, Stagg HR, *et al.* Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. *BMC Med* 2016; 14: 21.
- 14 Walker TM, Kohl TA, Omar SV, *et al.* Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect Dis* 2015; 15: 1193–1202.
- 15 Merker M, Blin C, Mona S, *et al.* Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. *Nat Genet* 2015; 47: 242–249.
- 16 Kohl TA, Diel R, Harmsen D, *et al.* Whole-genome-based *Mycobacterium tuberculosis* surveillance: a standardized, portable, and expandable approach. *J Clin Microbiol* 2014; 52: 2479–2486.
- 17 Zignol M, Cabibbe AM, Dean AS, *et al.* Genetic sequencing for surveillance of drug resistance in tuberculosis in highly endemic countries: a multi-country population-based surveillance study. *Lancet Infect Dis* 2018; 18: 675–683.
- 18 Miotto P, Tessema B, Tagliani E, *et al.* A standardised method for interpreting the association between mutations and phenotypic drug resistance in *Mycobacterium tuberculosis*. *Eur Respir J* 2017; 50: 1701354.
- 19 Tortoli E, Fedrizzi T, Meehan CJ, *et al.* The new phylogeny of the genus *Mycobacterium*: the old and the news. *Infect Genet Evol* 2017; 56: 19–25.
- 20 Merker M, Kohl TA, Roetzler A, *et al.* Whole genome sequencing reveals complex evolution patterns of multidrug-resistant *Mycobacterium tuberculosis* Beijing strains in patients. *PLoS One* 2013; 8: e82551.
- 21 Pankhurst LJ, Del Ojo Elias C, Votintseva AA, *et al.* Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. *Lancet Respir Med* 2016; 4: 49–58.
- 22 Cabibbe AM, Trovato A, De Filippo MR, *et al.* Countrywide implementation of whole genome sequencing: an opportunity to improve tuberculosis management, surveillance and contact tracing in low incidence countries. *Eur Respir J* 2018; 51: 1800387.
- 23 Shea J, Halse TA, Lapierre P, *et al.* Comprehensive whole-genome sequencing and reporting of drug resistance profiles on clinical cases of *Mycobacterium tuberculosis* in New York State. *J Clin Microbiol* 2017; 55: 1871–1882.
- 24 Marais BJ, Walker TM, Cirillo DM, *et al.* Aiming for zero tuberculosis transmission in low-burden countries. *Lancet Respir Med* 2017; 5: 846–848.
- 25 Tagliani E, Cirillo DM, Ködmön C, *et al.* EUSeqMyTB to set standards and build capacity for whole genome sequencing for tuberculosis in the EU. *Lancet Infect Dis* 2018; 18: 377.
- 26 Votintseva AA, Bradley P, Pankhurst L, *et al.* Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples. *J Clin Microbiol* 2017; 55: 1285–1298.
- 27 McNerney R, Clark TG, Campino S, *et al.* Removing the bottleneck in whole genome sequencing of *Mycobacterium tuberculosis* for rapid drug resistance analysis: a call to action. *Int J Infect Dis* 2017; 56: 130–135.
- 28 Global Laboratory Initiative (GLI), Stop TB Partnership. GLI Model TB Diagnostic Algorithms. 2018. [www.stoptb.org/wg/gli/assets/documents/GLI\\_algorithms.pdf](http://www.stoptb.org/wg/gli/assets/documents/GLI_algorithms.pdf)
- 29 Schleusener V, Köser CU, Beckert P, *et al.* *Mycobacterium tuberculosis* resistance prediction and lineage classification from genome sequencing: comparison of automated analysis tools. *Sci Rep* 2017; 7: 46327.
- 30 Crisan A, McKee G, Munzner T, *et al.* Evidence-based design and evaluation of a whole genome sequencing clinical report for the reference microbiology laboratory. *PeerJ* 2018; 6: e4218.
- 31 Lee RS, Pai M. Real-time sequencing of *Mycobacterium tuberculosis*: are we there yet? *J Clin Microbiol* 2017; 55: 1249–1254.