**Type of paper:** Narrative Review

**Title:** *Mycobacterium tuberculosis* and whole genome sequencing: a practical guide and online tools available for the clinical microbiologist

**Authors:** Giovanni Satta[1,2], Alessandro Atzeni[1], Timothy D. McHugh[1]

1. Centre for Clinical Microbiology, Department of Infection, University College London, UK

2. Imperial College Healthcare NHS Trust, London, UK

**Corresponding author:**
Dr Giovanni Satta
Research Fellow - University College London, Centre for Clinical Microbiology, Department of Infection
Consultant in Medical Microbiology and Infectious Diseases - Imperial College Healthcare NHS Trust, London
Giovanni.satta@nhs.net

**Keywords:** *Mycobacterium tuberculosis*, Whole Genome Sequencing, analysis

**Abstract**

Whole genome sequencing (WGS) has the potential to revolutionize the diagnosis of *Mycobacterium tuberculosis* (MTB) but the lack of bioinformatic expertise among clinical microbiologists is a barrier for adoption. Software products for analysis should be simple, free of charge, able to accept data directly from the sequencer (FASTQ files) and to provide the basic functionalities all-in-one.

The main aim of this narrative review is to provide a practical guide for the clinical microbiologist, with little or no practical experience of WGS analysis, with a specific focus on software products tailored made for MTB analysis.

With sequencing performed by an external provider, it is now feasible to implement WGS analysis in the routine clinical practice of any microbiology laboratory, with the potential to detect resistances weeks before traditional phenotypic culture methods, but the clinical microbiologist should be aware of some major limitations.

## Introduction

Tuberculosis (TB) still remains a global health problem[1]. Identification and susceptibility testing with conventional culture methods can take up to eight weeks for the final report[2], whilst targeted molecular tests (e.g. Xpert MTB/RIF and other PCR based methods), although much more rapid, only examine a limited number of target regions[3]. Whole genome sequencing (WGS) has been shown to provide a rapid and comprehensive view of the genotype of *Mycobacterium tuberculosis* (MTB), with the potential to reliably predict drug susceptibility within a clinically relevant timeframe[4]. In addition, it provides the highest resolution when investigating outbreaks[5].

However, the lack of bioinformatic expertise among clinical microbiologists is a potential barrier for clinical adoption. The Wellcome Trust Sanger Institute (UK)[6], the Galaxy project[7] (Pennsylvania State and John Hopkins universities - USA) and the Broad Institute[8] (Massachusetts Institute of Technology and Harvard University - USA), all combined offer hundreds of different software products for the analysis of WGS data but their high number and various functionalities often leave the unexperienced user confused in front of such enormous choice. It is evident that the use of WGS, despite its great potential, may be hampered by the complexity of data and its analysis. Hence, there is a need for simpler software tools, free of charge, able to analyze data directly from the initial FASTQ files and to provide the basic functionalities all-in-one.

Thus, the main aim of this narrative review is to provide a guide for the clinical microbiologist, with little or no practical experience in WGS analysis, with a specific focus on software products tailored made for MTB analysis. It is important to keep in mind that this review is not aimed to be a comprehensive, systematic description of all next generation sequencing platforms and analytic tools, but we have made the assumption that the sequencing is performed by an external provider and a FASTQ file is the end result received.

## Extraction of genomic DNA and sequencing formats

Extraction of genomic DNA from mycobacteria still requires special consideration because of the reduced bacterial load and the difficulty in extraction[9]. Minimum concentration of genomic DNA required is generally 10ng/µl with a 260/280 ratio of at least 1.8 (current requirement at Public Health England – PHE), but lower concentration may be accepted depending on the library preparation used/provider[10]. We continue to use the cetyltrimethylammonium bromide

(CTAB) method[11,12,13] and this is best performed from Löwenstein–Jensen culture. Other methods are available, including commercial extraction kits[14,15]. Once the DNA has been extracted, quantification needs to be confirmed using a Qubit fluorometer (Thermo Fisher Scientific, USA) or other alternative method[16] and the sample is now ready to be sent to the external provider for the sequencing.

All output files should be easily downloadable using a FTP (File Transfer Protocol) server (e.g. FileZilla software, freely available online under General Public License). Different format files[17] are used (Table 1) and they represent another major limitation in the analysis of WGS data. In addition, knowledge of Linux operating system and command line[18] is often essential for most of the advanced bio-informatics analysis.

**Free online tools available for WGS analysis in MTB**

We have reviewed four software programs focused on the rapid identification and susceptibility testing of MTB using WGS technology. As already mentioned, there are numerous software products available on the market but only programs available online (or downloadable in one click), free of charge, in English language and able to directly analyze raw FASTQ files (without any knowledge of Linux command line) were included. An overview of their different functionalities is shown in Tables 2 and 3.

- *Mykrobe predictor*[19] is a software package developed at the Wellcome Trust Centre for Human Genetics and collaborators[20]. The software presents a user friendly graphic interface in which raw sequence data in FASTQ format can be uploaded to generate a fast report (within 3 minutes), easily interpretable by clinicians. Files can be uploaded singularly (upload of paired end reads and multiple data is not allowed) and resistance panels and phylogenetic lineages are detected. The main limitation is that it does not provide information about sample quality, phylogenetic tree and linkage networks (Table 2). It detects the majority of first and second line drugs (Table 3). It is the only product that can be downloaded on your desktop and used without an internet connection.
- *TB Profiler*[21] was developed at the London School of Hygiene & Tropical Medicine and it processes raw sequence data identifying strain type and known drug resistance markers (with a drug resistance library containing 1,325 mutations) [22]. Sequence data can be submitted as single end or paired end read (FASTQ file) and processing time is under 10 minutes per sample plus queuing time (variable depending on concurrent demand). The displayed output provides information about resistances to 11 TB drugs (Table 3) and

lineages specific mutations. However, phylogenetic tree and linkage networks are not available (Table 2).

- *TGS-TB*[23]**,** Total Genotyping Solution for *Mycobacterium tuberculosis* (TGS-TB) is an *all-in-one* web based tool developed by Sekizuka et al[24]. Information about number of trimmed map reads and coverage region depth is provided. The resistances target list permits detection of genetic alterations for the majority of first and second line drugs (Table 3). In addition, up to 10 paired-ends FASTQ files can be uploaded simultaneously and the comparison of samples with phylogenetic tree allows outbreak investigation.
- *PhyRes SE*[25]**,** the Phylo-Resistance Search Engine (PhyResSE)[26] is designed to enable nonspecialized users to extract phylogenetic and resistance information from WGS data. Single end or paired end reads can be freely submitted and multiple file selection for upload is supported. A validation process runs to reject improperly formatted data files. Rigorous preprocessing and QC at the level of raw data, mapping performance, and individual SNPs are unique characteristics of the system. After preprocessing, variants are called and provided in VCF format and as HTML table carrying additional information about amino acid changes and what is known in terms of association with genotype or resistance. The majority of first and second line drugs are detected (Table 3).

**Discussion**

WGS holds great potential for the rapid diagnosis of drug-resistant TB and it does offer some added value when compared to traditional susceptibility testing. Phenotypic methods for TB are expensive, technically complex and time consuming[27]. It is current practice to initially test only for first line antibiotics (Rifampicin, Isoniazid, Ethambutol, Pyrazinamide, Streptomycin) and to subsequently perform second and third line susceptibility testing in case of drug resistance. This, combined with possible sample contamination or failure to grow in specific testing media, often causes delays of weeks (and even months) before receiving the final susceptibility profile. Full WGS diagnostics could theoretically be generated within 9 days (weeks in advance comparing to culture) and it could also be 7% cheaper than the present diagnostic workflows[28].

However, a recent report[29] from the EUCAST subcommittee on the role of WGS in antimicrobial susceptibility testing (AST) of bacteria, concluded that available published evidence is currently insufficient and it does not support the use of WGS inferred susceptibility to guide clinical decision making. It has also highlighted some specific issues regarding MTB, in particular the current discrepancies between genotype and phenotype and the need of large datasets to

clarify the role of rare resistance mechanisms and the level of resistance conferred by different mutations. Also, WGS does not seem to resolve some of the current problems of traditional susceptibility testing, in particular Pyrazinamide sensitivity[30], and it may actually over-report fluoroquinolones resistance in case of rare mutations conferring only higher MIC[31] or miss hetero-resistance without deep sequencing[32].

Accreditation (ISO 15189 and others) is another possible problem. None of the mentioned software tools is actually licensed for clinical/diagnostic use. One exception is the analysis infrastructure that is currently being evaluated by Public Health England in Birmingham (UK)[33].

WGS clearly has a role in public health interventions and in the detection of outbreaks and transmission events. Several studies confirm this role[34,35,] and the higher resolution compared to MIRU-VNTR typing, IS6110 RFLP typing and spoligotyping methods. Not all the online tools considered will be useful for this, with only TGS-TB and PhyResSE allowing phylogenetic trees and with TGS-TB also performing more in depth network analysis (including SNPs detection). However, several challenges remain before WGS can be routinely used in outbreak investigation and clinical practice[36], from better integration with conventional epidemiology and healthcare informatics systems to the need of moving beyond single-nucleotide polymorphisms (SNPs) typing to embrace full range of genome variation, including within-patient variation.

Unfortunately, many practical questions (how to deal with phenotype/genotype discrepant results, mutations conferring low level of resistance and hetero-resistance) are still unanswered and, based on these considerations, WGS is unlikely to completely replace phenotypic AST for TB in the near future. The clinical microbiologist should be aware of these limitations but this should not prevent the use of WGS.

This review has summarized the current online software tools available to support the clinical microbiologist in the analysis of WGS data in MTB. It is feasible to implement this approach in the routine clinical practice of any microbiology laboratory (with sequencing performed by an external provider). Not only would this inform individual patient management but could contribute to production and comparison of large independent datasets. This is essential if we are to confirm the perceived utility of WGS for rapid susceptibility testing in MTB.

## Tables and figures

| File formats used in Whole Genome Sequencing analysis | |
|---|---|
| ***FASTQ*** | FASTQ is the original and predominant file format when receiving WGS data from the sequencing provider (or in house analysis using an Illumina platform – Illumina, San Diego, USA) and all other formats will have to be created from it using the numerous software tools available on the market. The FASTA file is a similar version but without the quality metrics. |
| **SAM/BAM** | A SAM (Sequence Alignment Map) file represents a generic format for storing large nucleotide sequence alignments. It was created with the objective to obtain a well-defined interface between alignment and downstream analyses, including variant detection, genotyping and assembly. A BAM file is the binary version (or simply a lighter version) and it is easier to be handled with common computers. |
| **VCF** | This Variant Call Format was developed by the 1000 Genomes project to encode SNPs and other structural genetic variants. It allows representation of a wide variety of genomic variation with respect to a single reference sequence. |
| ***SFF*** | Standard Flowgram Format is the equivalent of FASTQ but from other platforms other than Illumina. |
| **Many other files are available (i.e. SRA, GFF, GTF, BED) but they are used for more specialized analysis** | |

**Table 1:** Most common files used for WGS output formats

| Functionality | Mykrobe predictor | TB Profiler | TGS-TB | PhyResSE |
|---|---|---|---|---|
| **Resistance** | ✓ | ✓ | ✓ | ✓ |
| **Lineage** | ✓ | ✓ | ✓ | ✓ |
| **Phylogenetic tree** | × | × | ✓ | ✓ |
| **Network analysis** | × | × | ✓ | × |
| **Quality Control** | × | × | ✓ | ✓ |
| **Time: per sample of 100 Mb** | 3 min | 5 min | 10 min | 15 min |

**Table 2:** Comparison of the functionalities offered by web based software tools for the analysis of whole genome sequencing data of *Mycobacterium tuberculosis*.

| Drug | Mykrobe Predictor | TB Profiler | TGS-TB | PhyResSE |
|---|---|---|---|---|
| **Isoniazid (INH)** | ✓ | ✓ | ✓ | ✓ |
| **Rifampicin (RIF)** | ✓ | ✓ | ✓ | ✓ |
| **Pyrazinamide (PZA)** | × | ✓ | ✓ | ✓ |
| **Ethambutol (EMB)** | ✓ | ✓ | ✓ | ✓ |
| **Streptomycin (SM)** | ✓ | ✓ | ✓ | ✓ |
| | | | | |
| **Fluoroquinolones (FQs)** | ✓ | ✓ | ✓ | ✓ |
| **Kanamycin (KAN)** | ✓ | ✓ | ✓ | ✓ |
| **Amikacin (AMK)** | ✓ | ✓ | ✓ | ✓ |
| **Capreomycin (CAP)** | ✓ | ✓ | ✓ | ✓ |
| **Ethionamide (ETH)** | × | ✓ | ✓ | ✓ |

| | | | | |
|---|---|---|---|---|
| **p-Amino salicylic acid (PAS)** | × | ✓ | × | × |
| **Cycloserine (CS)** | × | × | × | × |

**Table 3:** Comparison of first and second line anti-tuberculous drugs resistance detection offered by web based software tools.

## Transparency declaration

The Authors declare that no external funding was received for writing this paper and that they have no competing interests.

## Contributions

GS, AA, TDM conceived, designed the study and wrote the paper. All authors read and approved the final manuscript.

---

**References**

[1] World Health Organization website, last accessed on 1st July 2016 at: http://www.who.int/tb/publications/global_report/en/

[2] Pfyffer GE, Wittwer F. Incubation time of mycobacterial cultures: how long is long enough to issue a final negative report to the clinician? J Clin Microbiol. 2012 Dec;50(12):4188-9.

[3] Boehme, C. C. et al. Rapid molecular detection of tuberculosis and rifampin resistance. N Engl J Med 2010; 11, 1005–1015.

[4] Witney AA, Cosgrove CA, Arnold A, Hinds J, Stoker NG, Butcher PD. Clinical use of whole genome sequencing for Mycobacterium tuberculosis. BMC Med. 2016 Mar 23;14:46.

[5]Walker TM, Ip CL, Harrell RH et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis. 2013 Feb;13(2):137-46.

[6] Wellcome Trust Sanger Institute website, last accessed on 28th June 2016 at: http://www.sanger.ac.uk/science/tools/categories/analysis/all

[7] The Galaxy project website, last accessed on 28th June 2016 at: https://galaxyproject.org/

[8] Broad Institute website, last accessed on 28th June 2015 at: http://www.broadinstitute.org/scientific-community/software

[9] Käser M, Ruf MT, Hauser J, Pluschke G. Optimized DNA preparation from mycobacteria. Cold Spring Harb Protoc. 2010 Apr;2010(4):pdb.prot5408.

[10] Broad Institute website, Sample Quality and Quantiy Specifications, last accessed on 23rd August 2016, available at: https://www.broadinstitute.org/scientific-community/science/platforms/genomics/sample-quality-and-quantity-specifications

[11] Belisle JT, Mahaffey SB, Hill PJ. Isolation of mycobacterium species genomic DNA. Methods Mol Biol 2009; 465:1- 12.

[12] Kent L, McHugh TD, Billington O, Dale JW, Gillespie SH. Demonstration of homology between IS6110 of Mycobacterium tuberculosis and DNAs of other Mycobacterium spp.? J Clin Microbiol. 1995;33(9):2290-3.

[13] van Soolingen D, Hermans PW, de Haas PE, Soll DR, van Embden JD. Occurrence and stability of insertion sequences in Mycobacterium tuberculosis complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. J Clin Microbiol. 1991;29:2578–2586.

[14] Votintseva AA, Pankhurst LJ, Anson LW, Morgan MR, Gascoyne-Binzi D, Walker TM, Quan TP, Wyllie DH, Del Ojo Elias C, Wilcox M, Walker A, Peto TE, Crook DW. Mycobacterial DNA extraction for whole-genome sequencing from early positive liquid (MGIT) cultures. J Clin Microbiol 2015 Apr;53(4):1137-43.

[15] Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZ, Depledge DP, Nikolayevskyy V, Broda A, Stone MJ, Christiansen MT, Williams R, McAndrew MB, Tutill H, Brown J, Melzer M, Rosmarin C, McHugh TD, Shorten RJ, Drobniewski F, Speight G, Breuer J. Rapid Whole-Genome Sequencing of Mycobacterium tuberculosis Isolates Directly from Clinical Samples. J Clin Microbiol. 2015 Jul;53(7):2230-7.

[16] Robin JD, Ludlow AT, LaRanger R, Wright WE, Shay JW. Comparison of DNA Quantification Methods for Next Generation Sequencing. Sci Rep. 2016 Apr 6;6:24067.
[17] University of Wisconsin – Madison, Biotechnology Center website, last accessed 13th July 2016 at: https://www.biotech.wisc.edu/services/dnaseq/sequencing/Ion_Torrent_PGM/file_formats
[18] Linux website, last accessed 1st July 2016 at: https://www.linux.com/what-is-linux
[19] Mykrobe predictor website, last accessed on 3rd July 2016 at: http://www.mykrobe.com/products/predictor/#tb
[20] Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, Earle S, Pankhurst LJ, Anson L, de Cesare M, Piazza P, Votintseva AA, Golubchik T, Wilson DJ, Wyllie DH, Diel R, Niemann S, Feuerriegel S, Kohl TA, Ismail N, Omar SV, Smith EG, Buck D, McVean G, Walker AS, Peto TE, Crook DW, Iqbal Z.Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. Nat Commun. 2015 Dec 21;6:10063.
[21] TB Profiler website, last accessed on 3rd July 2016 at: http://tbdr.lshtm.ac.uk
[22] Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, Mallard K, Nair M, Miranda A, Alves A, Perdigão J, Viveiros M, Portugal I, Hasan Z, Hasan R, Glynn JR, Martin N, Pain A, Clark TG. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med. 2015 May 27;7(1):51.
[23] Total Genotyping Solution for Mycobacterium tuberculosis (TGS-TB) website, last accessed on 3rd July 2016 at: https://gph.niid.go.jp/tgs-tb/index_tb.html
[24] Sekizuka T, Yamashita A, Murase Y, Iwamoto T, Mitarai S, Kato S, Kuroda M. TGS-TB: Total Genotyping Solution for Mycobacterium tuberculosis Using Short-Read Whole-Genome Sequencing. PLoS ONE 2015; 10(11): e0142951.
[25] Phylo-Resistance Search Engine (PhyResSE) website, last accessed on 3rd July 2016 at: http://phyresse.org
[26] Feuerriegel S, Schleusener V, Beckert P, Kohl TA, Miotto P, Cirillo DM, Cabibbe AM, Niemann S, Fellenberg K. PhyResSE: web tool delineating Mycobacterium tuberculosis antibiotic resistance and lineage from whole-genome sequencing data. J Clin Microbiol 2015 Jun;53(6):1908-14.
[27] Koser CU, Ellington MJ, Cartwright EJ et al. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog 2012; 8: e1002824.
[28] 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.
[29]EUCAST website, last accessed on 3rd July 2016 at : http://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/Consultation/2016/EUCAST_WGS_report_consultation_20160511.pdf
[30] Javid B, Török ME. Whole-genome sequencing for the diagnosis of drug-resistant tuberculosis. Lancet Infect Dis. 2016 Jan;16(1):17
[31] Malik S, Willby M, Sikes D, Tsodikov OV, Posey JE. New insights into fluoroquinolone resistance in Mycobacterium tuberculosis: functional genetic analysis of gyrA and gyrB mutations. PLoS One. 2012;7(6):e39754.
[32] Eilertson B, Maruri F, Blackman A et al. High proportion of heteroresistance in gyrA and gyrB in fluoroquinolone-resistant Mycobacterium tuberculosis clinical isolates. Antimicrob Agents Chemother 2014; 58:3270-5.
[33] Robinson E, Bawa Z, Smith EG. Diagnosis of tuberculosis in England in the molecular genomic era. The Bulletin of The Royal College of Pathologists 2016; 175: 163-166.
[34] Gardy JL, Johnston JC, Sui SJH, Cook VJ, Shah L, Brodkin E, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med. 2011;364(8):730–9.
[35] Török ME, Reuter S, Bryant J, Köser CU, Stinchcombe SV, Nazareth B, et al. Rapid whole-genome sequencing for investigation of a suspected tuberculosis outbreak. J Clin Microbiol. 2013;51(2):611–4.
[36] Robinson ER, Walker TM, Pallen MJ. Genomics and outbreak investigation: from sequence to consequence. Genome Med. 2013 Apr 29;5(4):36.