ERJ Methods

Title: Whole Genome Sequencing of *Mycobacterium tuberculosis*

Andrea M Cabibbe[1], Timothy M Walker[2], Stefan Niemann[3], Daniela M Cirillo[1]

Affiliations: [1]Emerging Bacterial Pathogens Unit, IRCCS San Raffaele Scientific Institute, Milan, Italy; [2]Nuffield Department of Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK; [3]Molecular and Experimental Mycobacteriology, Research Center Borstel, Borstel, Germany

Correspondence: Daniela Maria 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

Next Generation Sequencing (NGS) technologies using massively parallel processing to interrogate pathogen genomes in days are revolutionizing the clinical microbiology practice [1]. Indeed, deep sequencing of selected genomic regions (targeted NGS) and whole genome sequencing (WGS) offer unprecedented resolution for genotyping, outbreak investigation and determination of known sequence variants involved in antimicrobial resistance (AMR). WGS-based approach has 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 *Mycobacteria* spp. (including *Mycobacterium tuberculosis*, MTB) [3].

Sequencing technologies

For years, the clinical microbiology laboratories relied on the use of conventional phenotypic and genotypic methods for species identification, detection of antibiotic resistance, and for studying transmission of bacteria responsible of hospital- or community-acquired infections that, although somewhat successful, have long turnaround time (TAT), specific infrastructure requirements, low discriminatory power to distinguish highly genetically-related strains, and technological constraints preventing to consider a wider range of genomic loci simultaneously [4]. On the other hand, the use of Sanger DNA sequencing since the 1970s enabled to reach important milestones in bacterial sequencing (e.g. first gene/genome sequences) but was still limited for a broad application due to the sophisticated requirements, complexity and high costs for sequencing of extended genomic regions [5]. The advent of pyrosequencing, but mainly of NGS methods in mid-2000s and their next expansion increase tremendously the sequencing output at lower costs compared to Sanger [6]. The NGS platforms generate short (50-400 bp) or long (1-100 kb) reads (i.e. sequence of a unique DNA fragments obtained at the end of the sequencing reaction), producing billions of sequences per run with different chemistry, output, accuracy and costs [5]. The workflow is similar for all the technologies: extraction of high molecular weight DNA; preparation of libraries (i.e. collection of DNA fragments) through DNA fragmentation (enzymatic or mechanical), barcoding and PCR amplification; clustering (clonal amplification of single DNA fragments); automated single- or paired-end (i.e. sequencer reads both ends of a DNA fragment) sequencing [5]. Among the major players in the field, Illumina Inc. (San Diego, CA, USA) uses the sequencing by synthesis (bridge PCR) chemistry, and offers a wide range of “benchtop” (MiniSeq, MiSeq) or high-throughput (NextSeq 500, HiSeq 2500, Novaseq) solutions allowing higher output at lower costs for generation of high-quality short reads. Thermo Fisher Scientific Inc. (Waltham, MA, USA) “benchtop” (PGM, 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 (Pacific Biosciences of California Inc., Menlo Park, CA, USA: RSII, Sequel; Oxford Nanopore Technologies Limited, Oxford, UK: MinION) take advantage of single-molecule real-time chemistry to generate long reads suitable particularly for *de novo* assembly in absence of reference genomes, and sequencing of complex regions as the repetitive ones. Throughput and quality are 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 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 AMR [9]. Multidrug (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% of susceptible forms) [9]. Phenotypic testing is the standard for drug susceptibility testing (DST) but, due to the slow growth rate of MTB, high-level biosafety infrastructure required, poor reproducibility and uncertainties around the proposed critical concentrations for some drugs, remains challenging particularly in resource-limited settings [10]. As MTB resistance emerges from single nucleotide polymorphisms (SNPs), insertions and deletions in genes coding for drug targets or involved in drug metabolic pathways, efflux pumps and compensatory mechanisms, methods detecting DNA mutations represent a breakthrough for the rapid, simple and standardized management of DR-TB cases [10]. There's 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, capable of reaching also the peripheral TB laboratories [11]. However, these tools target only the "hot-spot" regions of few genes to detect resistance to a restricted number of drugs, and not always informative of the exact nucleotide change leading to different phenotypic expression.

All-in-one solutions allowing to guide individualized clinical decisions also for the most complicated resistant cases are needed, at least at reference level, and NGS is taking a leading role in this regard [11]. The absence of integrative vectors and low mutation rate make the MTB genome well suited for sequencing, although the presence of repetitive and hard-to-sequence regions (high GC content) requires enough depth of genome-wide sequence coverage, with important cost implications for the NGS platforms [10].

Many studies explored the use of WGS for TB resistance prediction, transmission dynamics and population structure of MTB complex [10, 12, 13]. Genome sequencing is currently used for fast detection of the whole spectrum of resistance [14] and genotyping [15], and for public health-related topics as epidemiological [16] and drug resistance [17] surveillance. Other-published studies describe the: discovery of targets involved in phenotypic resistance and large-scale validation of resistance-conferring mutations [18]; taxonomy [19]; characterization of hetero-resistance and mixed infections; 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], as in such context the fast tracking of DR-TB and transmission events would help to reach TB elimination shortly [24, 25]. The results obtained using mainly Illumina technology showed the effectiveness of WGS from culture for routine diagnostic workflow, in terms of accuracy, time (reports available several days earlier than phenotypic DST) and costs. However, the on-going need of culturing poses a challenge to the full implementation of NGS as

an effective alternative to conventional methods (e.g. MTB/RIF Xpert, Cepheid), particularly in resource-limited settings. Efforts are thus being made to develop protocols to apply 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 challenged by the low starting TB material and contamination of other genetic material (human and oral flora). Targeted approaches taking advantage of the selective amplification of phylogenetic and DR-related regions may represent a suitable alternative for direct sequencing [27] (Figure).

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 1st and 2nd line drugs in replacement to phenotypic tests performing 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 power compared to traditional molecular methods, through SNP-based or core genome multilocus sequence typing approaches [16].

Sequencing implementation

Several aspects are considered for NGS implementation in TB clinical laboratories, including: strategic planning; procurement; budgeting; sample referral system; standard operating procedures (SOP); quality assurance; data management (storage, analysis, interpretation and reporting); human resources (staff and training). Scale-up of sequencing laboratories requires adequate infrastructure: areas for sample preparation (DNA extraction from clinical isolates/specimens considering related biosafety); molecular biology environment (pre-, post-PCR; space for NGS instrument); power supply; environment conditions (e.g. controlled temperature, humidity, vibration); network and internet connections; computing capacity. Equipment and reagents required are platform-specific: it's 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/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 bioinformatics background is required to handle output data for storage and to run analytic pipelines. Large-scale validation of bioinformatic analyses is still needed and international consortia are putting huge effort in the standardization of post-sequencing analysis processes for a reliable interpretation of DR-TB profiles and epidemiological links. User-friendly solutions have been developed avoiding the need of specific bioinformatics skills, both freely- [28] 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 requires adequate and continuous training and mentoring on wet procedures (background in molecular biology) and on post-sequencing processing.

Next steps

For a full implementation into routine workflows, NGS methods need to enter validation and certification programmes. Assessment of performance, accuracy and reproducibility, quality control steps, quality thresholds (depth/breadth of genome coverage), use of standards and development of SOPs, impact on TAT and clinical management, are all components to evaluate in a microbiology laboratory introducing sequencing [1, 7]. The laboratory will undergo external proficiency testing programmes that are already implemented in TB also for molecular testing and under evaluation specifically for sequencing. Simple but comprehensive clinical reports are crucial to address clinicians to the best decisions for management of TB cases. Given the huge amount of data generated by NGS but, at the same time, the incomplete knowledge for the interpretation of molecular mechanisms of resistance and for running epidemiological analyses,

development of report forms is still ongoing [29]. A report should give at least information on sequencing quality, identification of mutations to infer genotyping and DR profiles, providing details on the exact nucleotide changes and standardized prediction of resistance levels (ideally, based on a literature review of MIC 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, 30]. Additional work is needed to address the feasibility of WGS from clinical specimens, to standardize and automatize the laboratory procedures and post-sequencing analyses, and to implement the NGS platforms in low-resource, high-burden settings.

References

1. Rossen JWA, Friedrich AW, Moran-Gilad J, (ESGMD) ESGfGaMD. Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. *Clin Microbiol Infect* 2018: 24(4): 355-360.
2. WHO. Global Priority List of Antibiotic-resistant Bacteria to Guide Research, Discovery, and Development of New Antibiotics: World Health Organization; 2017.
3. Ellington MJ, Ekelund O, Aarestrup FM, Canton R, Doumith M, Giske C, Grundman H, Hasman H, Holden MTG, Hopkins KL, Iredell J, Kahlmeter G, Köser CU, MacGowan A, Mevius D, Mulvey M, Naas T, Peto T, Rolain JM, Samuelsen Ø, Woodford N. The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. *Clin Microbiol Infect* 2017: 23(1): 2-22.
4. Quainoo S, Coolen JPM, van Hijum SAFT, Huynen MA, Melchers WJG, van Schaik W, Wertheim HFL. Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. *Clin Microbiol Rev* 2017: 30(4): 1015-1063.
5. Besser J, Carleton HA, Gerner-Smidt P, Lindsey RL, Trees E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clin Microbiol Infect* 2018: 24(4): 335-341.
6. Goldberg B, Sichtig H, Geyer C, Ledeboer N, Weinstock GM. Making the Leap from Research Laboratory to Clinic: Challenges and Opportunities for Next-Generation Sequencing in Infectious Disease Diagnostics. *MBio* 2015: 6(6): e01888-01815.
7. Kozyreva VK, Truong CL, Greninger AL, Crandall J, Mukhopadhyay R, Chaturvedi V. Validation and Implementation of Clinical Laboratory Improvements Act-Compliant Whole-Genome Sequencing in the Public Health Microbiology Laboratory. *J Clin Microbiol* 2017: 55(8): 2502-2520.
8. APHL. Next Generation Sequencing Implementation Guide: Association of Public Health Laboratories; 2016.
9. WHO. Global tuberculosis report 2017: World Health Organization; 2017.
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(5): 433-442.
11. Cabibbe AM, Sotgiu G, Izco S, Migliori GB. Genotypic and phenotypic M. *tuberculosis* resistance: guiding clinicians to prescribe the correct regimens. *Eur Respir J* 2017: 50(6).
12. Walker TM, Merker M, Kohl TA, Crook DW, Niemann S, Peto TE. Whole genome sequencing for M/XDR tuberculosis surveillance and for resistance testing. *Clin Microbiol Infect* 2017: 23(3): 161-166.
13. Hatherell HA, Colijn C, Stagg HR, Jackson C, Winter JR, Abubakar I. Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. *BMC Med* 2016: 14: 21.
14. Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, Bradley P, Iqbal Z, Feuerriegel S, Niehaus KE, Wilson DJ, Clifton DA, Kapatai G, Ip CLC, Bowden R, Drobniewski FA, Allix-Béguec C, Gaudin C, Parkhill J, Diel R, Supply P, Crook DW, Smith EG, Walker AS, Ismail N, Niemann S, Peto TEA, Group MMMMI. Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect Dis* 2015: 15(10): 1193-1202.

15. Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, Blum MG, Rüsch-Gerdes S, Mokrousov I, Aleksic E, Allix-Béguec C, Antierens A, Augustynowicz-Kopeć E, Ballif M, Barletta F, Beck HP, Barry CE, Bonnet M, Borroni E, Campos-Herrero I, Cirillo D, Cox H, Crowe S, Crudu V, Diel R, Drobniewski F, Fauville-Dufaux M, Gagneux S, Ghebremichael S, Hanekom M, Hoffner S, Jiao WW, Kalon S, Kohl TA, Kontsevaya I, Lillebæk T, Maeda S, Nikolayevskyy V, Rasmussen M, Rastogi N, Samper S, Sanchez-Padilla E, Savic B, Shamputa IC, Shen A, Sng LH, Stakenas P, Toit K, Varaine F, Vukovic D, Wahl C, Warren R, Supply P, Niemann S, Wirth T. Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. *Nat Genet* 2015: 47(3): 242-249.
16. Kohl TA, Diel R, Harmsen D, Rothgänger J, Walter KM, Merker M, Weniger T, Niemann S. Whole-genome-based Mycobacterium tuberculosis surveillance: a standardized, portable, and expandable approach. *J Clin Microbiol* 2014: 52(7): 2479-2486.
17. Zignol M, Cabibbe AM, Dean AS, Glaziou P, Alikhanova N, Ama C, Andres S, Barbova A, Borbe-Reyes A, Chin DP, Cirillo DM, Colvin C, Dadu A, Dreyer A, Driesen M, Gilpin C, Hasan R, Hasan Z, Hoffner S, Hussain A, Ismail N, Kamal SMM, Khanzada FM, Kimerling M, Kohl TA, Mansjö M, Miotto P, Mukadi YD, Mvusi L, Niemann S, Omar SV, Rigouts L, Schito M, Sela I, Seyfaddinova M, Skenders G, Skrahina A, Tahseen S, Wells WA, Zhurilo A, Weyer K, Floyd K, Raviglione MC. 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. Miotto P, Tessema B, Tagliani E, Chindelevitch L, Starks AM, Emerson C, Hanna D, Kim PS, Liwski R, Zignol M, Gilpin C, Niemann S, Denkinger CM, Fleming J, Warren RM, Crook D, Posey J, Gagneux S, Hoffner S, Rodrigues C, Comas I, Engelthaler DM, Murray M, Alland D, Rigouts L, Lange C, Dheda K, Hasan R, Ranganathan UDK, McNerney R, Ezewudo M, Cirillo DM, Schito M, Köser CU, Rodwell TC. A standardised method for interpreting the association between mutations and phenotypic drug resistance in Mycobacterium tuberculosis. *Eur Respir J* 2017: 50(6).
19. Tortoli E, Fedrizzi T, Meehan CJ, Trovato A, Grottola A, Giacobazzi E, Serpini GF, Tagliazucchi S, Fabio A, Bettua C, Bertorelli R, Frascaro F, De Sanctis V, Pecorari M, Jousson O, Segata N, Cirillo DM. The new phylogeny of the genus Mycobacterium: The old and the news. *Infect Genet Evol* 2017: 56: 19-25.
20. Merker M, Kohl TA, Roetzer A, Truebe L, Richter E, Rüsch-Gerdes S, Fattorini L, Oggioni MR, Cox H, Varaine F, Niemann S. Whole genome sequencing reveals complex evolution patterns of multidrug-resistant Mycobacterium tuberculosis Beijing strains in patients. *PLoS One* 2013: 8(12): e82551.
21. Pankhurst LJ, Del Ojo Elias C, Votintseva AA, Walker TM, Cole K, Davies J, Fermont JM, Gascoyne-Binzi DM, Kohl TA, Kong C, Lemaitre N, Niemann S, Paul J, Rogers TR, Roycroft E, Smith EG, Supply P, Tang P, Wilcox MH, Wordsworth S, Wyllie D, Xu L, Crook DW, Group C-TS. Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. *Lancet Respir Med* 2016: 4(1): 49-58.
22. Cabibbe AM, Trovato A, De Filippo MR, Ghodousi A, Rindi L, Garzelli C, Baretti S, Allodi G, Mannino R, Rossolini GM, Bartoloni A, Tortoli E, Cirillo DM. Countrywide implementation of whole genome sequencing: an opportunity to improve tuberculosis management, surveillance and contact tracing in low incidence countries. *Eur Respir J* 2018.
23. Shea J, Halse TA, Lapierre P, Shudt M, Kohlerschmidt D, Van Roey P, Limberger R, Taylor J, Escuyer V, Musser KA. 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(6): 1871-1882.
24. Marais BJ, Walker TM, Cirillo DM, Raviglione M, Abubakar I, van der Werf MJ, Boehme C, Niemann S, Castro KG, Zumla A, Sintchenko V, Crook DW. Aiming for zero tuberculosis transmission in low-burden countries. *Lancet Respir Med* 2017: 5(11): 846-848.
25. Tagliani E, Cirillo DM, Ködmön C, van der Werf MJ, Consortium E. EUSeqMyTB to set standards and build capacity for whole genome sequencing for tuberculosis in the EU. *Lancet Infect Dis* 2018: 18(4): 377.
26. Votintseva AA, Bradley P, Pankhurst L, Del Ojo Elias C, Loose M, Nilgiriwala K, Chatterjee A, Smith EG, Sanderson N, Walker TM, Morgan MR, Wyllie DH, Walker AS, Peto TEA, Crook DW, Iqbal Z. Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. *J Clin Microbiol* 2017: 55(5): 1285-1298.

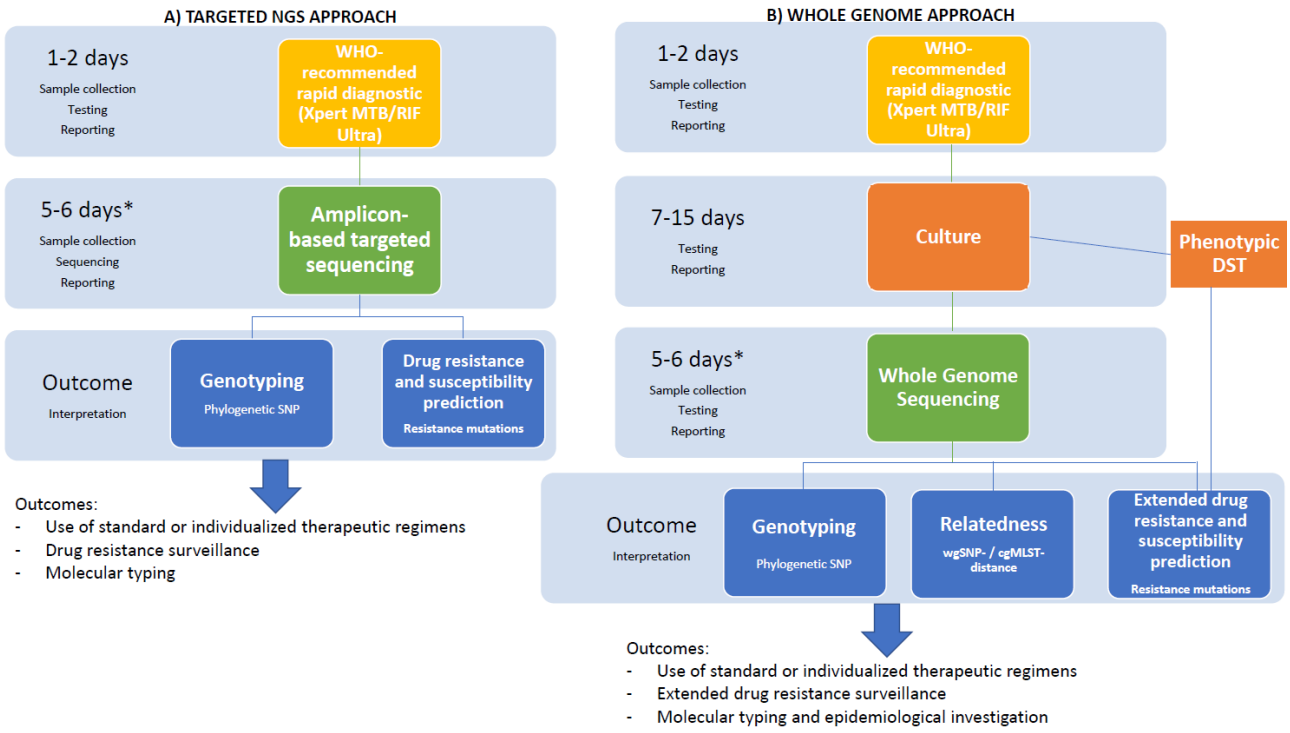
27. McNerney R, Clark TG, Campino S, Rodrigues C, Dolinger D, Smith L, Cabibbe AM, Dheda K, Schito M. Removing the bottleneck in whole genome sequencing of Mycobacterium tuberculosis for rapid drug resistance analysis: a call to action. *Int J Infect Dis* 2017: 56: 130-135.
28. Schleusener V, Köser CU, Beckert P, Niemann S, Feuerriegel S. Mycobacterium tuberculosis resistance prediction and lineage classification from genome sequencing: comparison of automated analysis tools. *Sci Rep* 2017: 7: 46327.
29. Crisan A, McKee G, Munzner T, Gardy JL. Evidence-based design and evaluation of a whole genome sequencing clinical report for the reference microbiology laboratory. *PeerJ* 2018: 6: e4218.
30. Lee RS, Pai M. Real-Time Sequencing of Mycobacterium tuberculosis: Are We There Yet? *J Clin Microbiol* 2017: 55(5): 1249-1254.

Figure: "Whole Genome Sequencing of *Mycobacterium tuberculosis*"

Two possible NGS-based scenarios and turnaround times to guide treatment decisions are presented. Modified GLI model TB diagnostic algorithms (Global Laboratory Initiative, March 2017) are outlined here and complemented with NGS approach. A) targeted NGS from clinical specimens (no culture required). B) whole genome sequencing from cultured samples. Xpert MTB/RIF (Ultra) is used in this illustrative flow as initial diagnostic test for persons being evaluated for TB.

1) If MTB detected / rifampicin resistance not detect, treat with first line regimen, and refer sample for:
A) tNGS that may promptly reveal additional resistance and guide appropriate treatment (e.g. for rifampicin-susceptible, isoniazid-resistant TB);
B) culture (and phenotypic DST). As soon as culture turns positive, perform WGS that may promptly reveal additional resistance and guide appropriate treatment (e.g. for rifampicin-susceptible, isoniazid-resistant TB), providing extended DST information

2) If MTB detected / rifampicin resistance detect, initiate treatment with second line regimen or shorter MDR-TB regimen, as appropriate, and refer sample for:
A) tNGS that may promptly reveal additional resistance and guide standardized or individualized MDR/RR-TB regimens;
B) culture (and phenotypic DST). As soon as culture turns positive, perform WGS that may promptly reveal additional resistance and guide standardized or individualized MDR/RR-TB regimens, providing extended DST information (including all repurposed and new drugs). WGS approach enables also high-resolution outbreak analysis to guide public health interventions.



(t)NGS: (targeted) Next Generation Sequencing; MTB: *M. tuberculosis*; DST: Drug Susceptibility Testing; MDR: Multidrug resistant; RR: rifampicin-resistant; SNP: Single Nucleotide Polimorphisms; wgSNP: whole genome Single Nucleotide Polimorphisms; cgMLST: core genome Multilocus Sequence Typing

*subjected to batching according to the NGS platform throughput