Whole genome sequencing of Mycobacterium tuberculosis: current standards and open issues

Conor J. Meehan, Galo A. Goig, Thomas A. Kohl, Lennert Verboven, Anzaan Dippenaar, Matthew Ezewudo, Maha R. Farhat, Jennifer L. Guthrie, Kris Laukens, Paolo Miotto, et al.

Whole genome sequencing of *Mycobacterium tuberculosis*: current standards and open issues

Conor J. Meehan[1], Galo A. Goig[2], Thomas A. Kohl[3,4], Lennert Verboven[5], Anzaan Dippenaar[6], Matthew Ezewudo[7], Maha R. Farhat[8,9], Jennifer L. Guthrie[10], Kris Laukens[11], Paolo Miotto[12], Boatema Ofori-Anyinam[13,14], Viola Dreyer[3,4], Philip Supply[15,16], Anita Suresh[17], Christian Utpatel[3,4], Dick van Soolingen[18], Yang Zhou[19], Philip M. Ashton[20], Daniela Brites[21,22], Andrea M. Cabibbe[12], Bouke C. de Jong[1], Margaretha de Vos[5], Fabrizio Menardo[21,22], Sebastien Gagneux[21,22], Qian Gao[23], Tim H. Heupink[5], Qingyun Liu[23], Chloé Loiseau[21,22], Leen Rigouts[1], Timothy C. Rodwell[17], Elisa Tagliani[12], Timothy M. Walker[24], Robin M. Warren[6], Yanlin Zhao[19], Matteo Zignol[25], Marco Schito[7], Jennifer Gardy[10], Daniela M. Cirillo[12], Stefan Niemann[3,4], Inaki Comas[2] and Annelies Van Rie[5]**

Abstract | Whole genome sequencing (WGS) of *Mycobacterium tuberculosis* has rapidly progressed from a research tool to a clinical application for the diagnosis and management of tuberculosis and in public health surveillance. This development has been facilitated by drastic drops in cost, advances in technology and concerted efforts to translate sequencing data into actionable information. There is, however, a risk that, in the absence of a consensus and international standards, the widespread use of WGS technology may result in data and processes that lack harmonization, comparability and validation. In this Review, we outline the current landscape of WGS pipelines and applications, and set out best practices for *M. tuberculosis* WGS, including standards for bioinformatics pipelines, curated repositories of resistance-causing variants, phylogenetic analyses, quality control and standardized reporting.

Mycobacterium tuberculosis complex (MTBC) pathogens are collectively the top infectious disease killer globally, causing an estimated 10 million new tuberculosis (TB) cases annually[1]. Increasingly, new TB cases are already resistant to rifampicin and isoniazid (termed 'multidrug-resistant TB'), the key first-line drugs[1]. Tackling the spread and drug resistance burden of *M. tuberculosis* requires concerted global effort in prevention, diagnosis, treatment and surveillance. Over the past decades, research and public health practices, including contact investigation and phenotypic methods for drug susceptibility testing (DST), have been complemented by molecular approaches. These can now provide rapid diagnosis, drug susceptibility profiling and an understanding of *M. tuberculosis* transmission dynamics[2,3].

Whole genome sequencing (WGS) approaches use DNA sequencing platforms to reconstruct the complete DNA sequence of an organism's genome. The small (~4.4Mb), single-chromosome genome of MTBC strains[4] is well suited to WGS approaches. Rapid, reliable and increasingly affordable WGS technologies can now guide all components of TB control: diagnosis, treatment, surveillance and source investigation[5,6] (FIG. 1). Individual strains of human and animal MTBC lineages can be identified by WGS[7–9], and drug resistance profiles can be predicted, especially well for first-line drugs[2], allowing prompt, appropriate initiation of treatment and the monitoring of the acquisition of drug resistance[10]. TB outbreaks can be identified with high resolution[11–13], including across borders[14,15], and disease control measures can be implemented. The analysis of the emergence, spread, genetic makeup and evolution of specific outbreak strains (for example, highly resistant or highly virulent clones) can allow the implementation of targeted measures[16–18].

WGS-based approaches are quickly moving from research laboratories to clinical care and public health applications. The WHO is already using WGS for drug resistance surveillance[19] and is scheduled to evaluate

Mycobacterium tuberculosis complex (MTBC). The genetically related group of organisms within the genus *Mycobacterium* that cause tuberculosis in humans or animals.

Drug susceptibility testing (DST). A procedure to determine if clinical isolates are resistant to antibiotics either by testing the inhibition in culture (phenotypic DST) or by identifying drug resistance-associated mutations (genotypic DST).

**e-mail: icomas@ibv.csic.es; annelies.vanrie@uantwerpen.be*

https://doi.org/10.1038/s41579-019-0214-5

Source investigation
The first case in a group of related individuals that transmitted the disease. Usually, identified during the development of an epidemiological investigation.

Löwenstein–Jensen
A selective culture solid medium commonly used to isolate *Mycobacterium tuberculosis* complex strains.

Mycobacteria Growth Indicator Tube
A tube that contains mycobacteria-selective culture liquid medium and is usually coupled to an automated instrument to read the results.

sequencing technologies for routine genotypic DST in 2019 (REF.[1]). As WGS-guided individualized treatment[20] and WGS-based surveillance systems[15] are being implemented in several countries (for example, the United Kingdom and the Netherlands) with more to come, accurate methods and standardized reporting are vital. At present, multiple WGS data analysis solutions exist that differ widely in scope, pipelines and output formats, with little standardization among them, making cross-comparisons and rigorous validation of these pipelines difficult. As clinical decisions such as the choice of a drug regimen for treatment may be influenced by differences in bioinformatic analyses, robustness of the pipeline used in clinically relevant prediction tools is crucial.

In this Review, we present the current state of the art for the three core MTBC WGS tasks: drug susceptibility profiling, transmission cluster detection and subspecies or lineage identification (strain typing). We highlight areas where general agreement in the analysis parameters or interpretation of the results has already been reached by the community. We also discuss areas where there is still open discussion about the best practices that will require more effort to reach a consensus in the future.

Author addresses

[1]Unit of Mycobacteriology, Department of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium.
[2]Institute of Biomedicine of Valencia, CSIC, Valencia, Spain.
[3]Molecular and Experimental Mycobacteriology, Priority Area Infections, Research Center Borstel, Borstel, Germany.
[4]German Center for Infection Research, Partner Site Hamburg–Lübeck–Borstel–Riems, Borstel, Germany.
[5]Tuberculosis Omics Research Consortium, Department of Epidemiology and Social Medicine, Institute of Global Health, Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium.
[6]DST–NRF Centre of Excellence for Biomedical Tuberculosis Research, South African Medical Research Council Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa.
[7]Critical Path Institute, Tucson, AZ, USA.
[8]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA.
[9]Pulmonary and Critical Care Medicine, Massachusetts General Hospital, Boston, MA, USA.
[10]University of British Columbia, Vancouver, Canada.
[11]Adrem Data Laboratory, Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium.
[12]Emerging Bacterial Pathogens Unit, Division of Immunology, Transplantation and Infectious Diseases, IRCCS San Raffaele Scientific Institute, Milan, Italy.
[13]Center for Global Health Security and Diplomacy, Ottawa, Canada.
[14]Food and Drugs Authority, Accra, Ghana.
[15]University of Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, Lille, France.
[16]U1019 — UMR 8204, Center for Infection and Immunity of Lille, Lille, France.
[17]Foundation for Innovative New Diagnostics, Geneva, Switzerland.
[18]National Tuberculosis Reference Laboratory, Centre for Infectious Disease Control, National Institute for Public Health and the Environment (RIVM), Bilthoven, Netherlands.
[19]National Center for Tuberculosis Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China.
[20]Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK.
[21]Swiss Tropical and Public Health Institute, Basel, Switzerland.
[22]University of Basel, Basel, Switzerland.
[23]Key Laboratory of Medical Molecular Virology, Ministry of Education and Health, School of Basic Medical Sciences, Fudan University, Shanghai, China.
[24]Nuffield Department of Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK.
[25]Global Tuberculosis Programme, World Health Organization, Geneva, Switzerland.

State of the art

The standard workflow for WGS analysis of MTBC strains (FIG. 2) involves culturing sputum specimens on solid (Löwenstein–Jensen) or liquid (Mycobacteria Growth Indicator Tube) media, extracting DNA from cells, library preparation and sequencing using short read technologies (for example, Illumina platforms)[21]. The complete MTBC WGS analysis pipeline involves several key steps, such as input data validation and quality control followed by mapping to a reference genome (often *M. tuberculosis* strain H37Rv) and detection of genomic variants such as SNPs and insertion or deletions (indels). Numerous resequencing pipelines for the MTBC currently exist, with currently no single gold standard. These pipelines typically exclude ~10% of the genome because erroneous mapping in certain regions results in false variant calls (PE and PPE gene families, other repetitive genes and mobile genetic elements)[4] and apply various criteria, such as read depth, base quality and strand bias, to filter out false positive variants. Finally, on the basis of the variants detected, several tasks can be performed, including (but not limited to) prediction of drug resistance and susceptibility profiles, strain typing and identification of transmission clusters.

Owing to the clonality of their genomes and their inability to undergo lateral gene transfer, MTBC strains acquire drug resistance primarily through variants in core genes or promoters[22,23]. Drug resistance and susceptibility profiles can be determined with high accuracy for many drugs used for the treatment of TB by comparing variant calls with lists of high-confidence resistance-conferring variants. These lists have been established primarily using genotype–phenotype associations identified from statistical analyses of large sets of clinical WGS data[24,25] (FIG. 3). A prime effort in the construction of these lists is the Relational Sequencing Tuberculosis Data Platform (ReSeqTB), where researchers from around the world can contribute data[26]. This database contains curated, aggregated genotypic and phenotypic information on global MTBC isolates accompanied by metadata including clinical outcome. Another important initiative is the Comprehensive Resistance Prediction for Tuberculosis: an International Consortium (CRyPTIC) project. CRyPTIC aims to better understand the relationship between genetic variants and minimum inhibitory concentrations for most drugs used for TB treatment[2]. By comparing the SNPs present in a sequenced isolate with these lists, WGS can predict not only resistance but also first-line pansusceptibility under specific conditions[2], replacing the need for phenotypic testing.

Similarly, strain classification of the seven major human-associated lineages, many of the animal-associated lineages and their sublineages can be derived directly from variant calls using lists of lineage-defining SNPs[7–9]. This is important for understanding population structure and potential phenotypic differences between lineages[27] and comparing isolates on the global level[18,28].

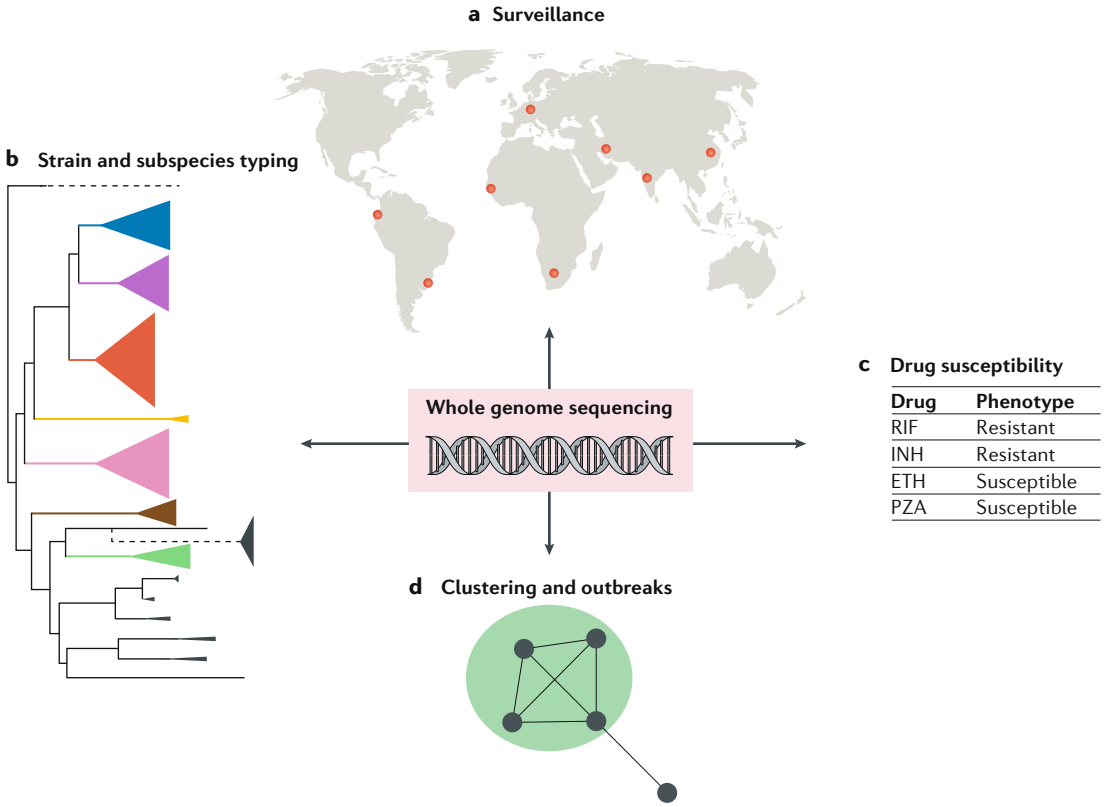


Fig. 1 | **Whole genome sequencing of *Mycobacterium tuberculosis*.** The primary applications for whole genome sequencing of *M. tuberculosis* in public health include international surveillance of prevalence and drug resistance (panel **a**), determination of the species or subspecies of *M. tuberculosis* complex isolates (panel **b**), determination of drug resistance patterns on the basis of the presence of specific SNPs (panel **c**) and identification of transmission clusters and outbreaks (panel **d**). ETH, ethambutol; INH, isoniazid; PZA, pyrazinamide; RIF, rifampicin. Panel **b** is adapted with permission from REF.[155], Springer Nature Limited.

The genomic data for a set of isolates can also be used for surveillance and transmission investigations. For this, the most common approach is to use a SNP cut-off-based clustering method, although genome-based multilocus sequence typing (MLST) has shown comparable results[29,30]. The SNP cut-off approach starts by constructing a list of high-confidence, unambiguous SNPs found in each isolate, often excluding indels and drug resistance-related sites. This filtering is important when predefined SNP distance thresholds are used to cluster strains and define recent transmission chains. Given the very low genetic diversity of the MTBC, thresholds of 5 or 12 SNPs are frequently used to suggest epidemiological links, although these thresholds were calibrated in low-incidence settings with a diverse strain population[31]. It is not yet clear if a single threshold can be used to detect epidemiologically linked cases in all timeframes and contexts. The MLST approach uses a predefined set of shared genes and assigns a number to each allele sequence identified for each gene. Coded allele combinations can be compared between strains to detect potential transmission clusters. Two schemes exist for this approach: core genome MLST (2,891 genes covering 2.86 million bases)[30] and an extended pangenome including 1,141 accessory loci (whole genome MLST)[11]. These WGS-based approaches have been shown to perform better than contact tracing and with higher resolution than classic approaches such as mycobacterial interspersed repetitive unit variable-number tandem repeat genotyping[12,13,29,30,32].

This currently recommended data processing workflow (FIG. 2) leading to SNP-based drug resistance profiling, transmission clustering at a given SNP cut-off and strain profiling using lineage-defining SNPs is often robust and reliable. However, steps towards standardization and validation of this workflow are required to ease integration into current clinical and public health initiatives.

Currently, two MTBC-specific pipelines are available that perform multiple core tasks in a single-installation set-up to produce genetic variant calls from raw Illumina sequence data (MTBseq[33] and UVP-ReSeqTB[34]). General WGS pipelines that are not specific to a certain pathogen can be used with an MTBC-specific reference genome and drug resistance database to achieve similar results[32,35–37]. Numerous custom-built pipelines also exist[8,38–45], often incorporating similar tools for mapping and variant calling, with additional accessory tools and in-house scripts to parse and refine outputs. A non-exhaustive list of such pipelines is given in Supplementary Table 1 to demonstrate the range of tools and settings that are routinely implemented. Lastly, pipelines specific for a single task, such as drug resistance prediction[24,46–50] or strain typing[7,49],

PE and PPE gene families
Families of genes that encode virulence factors in *Mycobacterium tuberculosis* complex strains. They have signature (proline)–proline–glutamate ((P)PE) motifs at their amino terminus.

Core genome MLST
A scheme that converts genome-wide SNP data into an allele-numbering system using a preselected set of core genes.

Whole genome MLST
A scheme that converts genome-wide SNP data into an allele-numbering system using a preselected set of core genes and additional accessory genes.

Contact tracing
The identification of possible contacts that interacted with an infected person (index case), often through questionnaires and interviews.

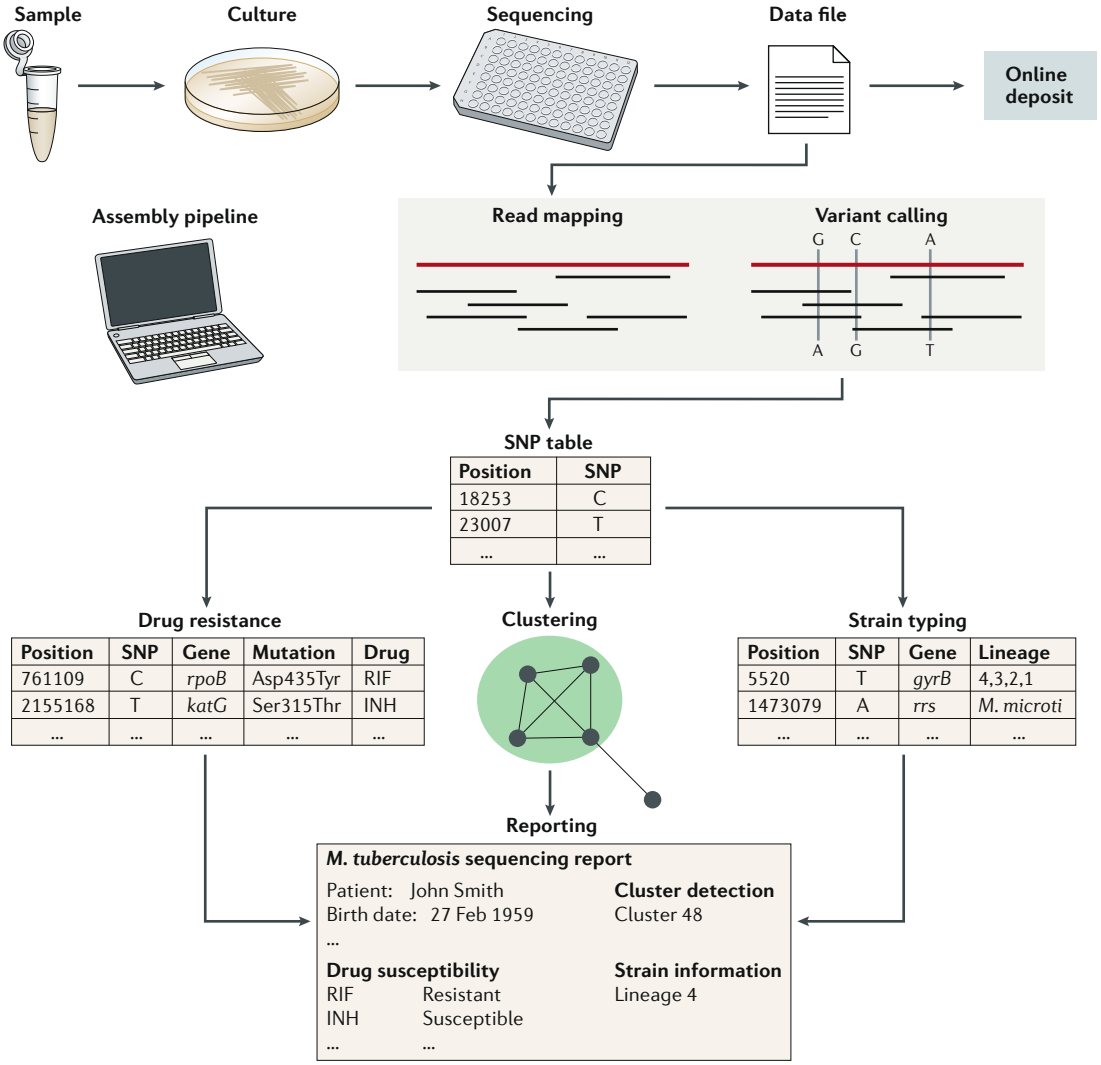


Fig. 2 | **Standard workflow for whole genome sequencing of *Mycobacterium tuberculosis* complex isolates.** A clinical sample (often sputum) is first cultured for up to 6 weeks, followed by genomic DNA extraction and sequencing. The resulting sequencing output (fastq data files) can be deposited online in public repositories and also run through standard SNP-calling pipelines. These pipelines first map reads to a reference genome (often *M. tuberculosis* strain H37Rv) and then call genomic variants, creating a table of SNPs. The resulting SNP lists can then be used for a variety of analyses, such as strain typing, transmission clustering and drug resistance profiling. The results of these tasks are then reported to the end user (for example, a clinician or researcher). INH, isoniazid; RIF, rifampicin.

are available and have been comprehensively compared elsewhere[51–54].

Validation and standardization

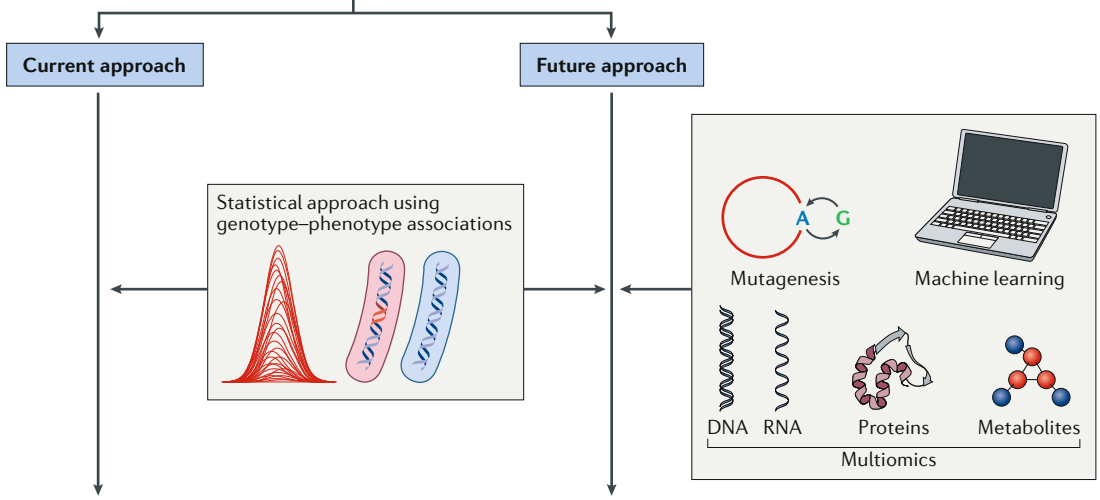
Before a workflow can become a gold standard, the validity of that workflow needs to be ensured for its intended uses. For MTBC WGS workflows, this essentially means ensuring virtually every variant that is reported is truly present in the isolate (validation) and each pipeline calls the same variants (standardization). Ideally, all steps of the workflow, from DNA extraction to sequencing, data analysis and reporting, should be standardized (or at least comparable) and well documented, and an external quality assessment programme should be in place. Efforts to standardize and validate pre-bioinformatics pipeline steps have been undertaken to great effect[21,53]. Pipeline standardization could be achieved through the use of a single pipeline in all settings or through validation with rigorous testing and convergence to a defined outcome for all pipelines developed. Since multiple pipelines have already been implemented (for example, MTBseq[33] for the EUSeqMyTB consortium and the Unified Variant Pipeline[34] for ReSeqTB) (Supplementary Table 1), agreement on validation criteria seems more realistic. Since WGS-based diagnostics present a potential paradigm shift for regulatory approvals, there is an urgent need to understand how to validate and standardize these multiple pipelines for clinical use[55]. In 2016, the US Food and Drug Administration released draft guidelines on sequencing-based infectious disease diagnostics, and bodies such as the WHO and the European Centre for Disease Prevention and Control are taking steps towards international standardizations of MTBC WGS[15,21,56].

Mycobacterial interspersed repetitive unit variable-number tandem repeat
Mycobacterium tuberculosis complex (MTBC)-specific variable tandem repeat locus used to genotype MTBC strains.

WGS pipelines
The bioinformatics section of the whole genome sequencing workflow, starting from raw sequencing files through to SNP calling and analyses.

Input whole genome sequencing data

	Strain 1	Strain 2	Strain 3	Strain 4
Lineage	4	3	1	*M. bovis*
***rpoB* mutation**	Ser450Leu	Ser450Leu	Ser450Leu	Gln429Ala
***pncA* mutation**	Val130Gly	Val130Gly	Arg123Gly	Gly108Ser
Phenotype	RIF resistant PZA resistant	RIF resistant PZA resistant	RIF resistant PZA susceptible	RIF susceptible PZA resistant



Resistance detected using a statistical approach

Gene	Mutation	Drug
rpoB	Ser450Leu	RIF
pncA	Val130Gly	PZA

Extended knowledge base

Gene	Mutation	Drug	Phenotype	Support
rpoB	Ser450Leu	RIF	Resistant	Statistical
rpoB	Gln429Ala	RIF	Susceptible	Mutagenesis
pncA	Val130Gly	PZA	Resistant	Statistical and mutagenesis
pncA	Arg123Gly	PZA	Susceptible	Machine learning
pncA	Gly108Ser	PZA	Resistant	Transcriptomics

Fig. 3 | **Current and potential approaches for determining resistance-related polymorphisms.** In the current approach, linked phenotypic–genotypic data derived from a variety of strains across the diversity of the *Mycobacterium tuberculosis* complex are passed through statistical approaches such as likelihood ratios to identify genetic variants that are likely related to drug resistance. In this example, the Ser450Leu mutation in *rpoB* is observed in phenotypically rifampicin (RIF)-resistant strains from multiple lineages. Thus, this mutation has a high probability of being associated with RIF resistance and is added to the list. The suggested future approach would complement this procedure with additional information from targeted mutagenesis, machine learning, multiomics and so on to detect drug resistance-causing SNPs that are too rare to be detected with a statistical approach only. For example, the Gly108Ser mutation in *pncA* is observed in only a single strain, but further confirmation of its association with pyrazinamide (PZA) resistance may be undertaken with other methods, allowing it to be added to the list. Additionally, such extensions could also determine variants that are definitely not associated with resistance (for example, *pncA* Arg123Gly).

Technical validation and external quality control of MTBC WGS. First, the extracted DNA needs to meet minimal standards as defined for a given WGS instrument[21]. Next, the pipeline to convert the raw sequencing reads into accurate variant calls should be technically valid, that is, call the correct variants. Although there is much debate about the reference standard to be used for technical validation of WGS pipelines, currently this is best undertaken by use of short read data sets derived from isolates with known complete genomes (for example, from long read sequencing)[57]. Mapping these read sets to their respective assembled genomes allows the rate of false positive and false negative SNPs called by the pipeline to be calculated. Ideally, to promote interoperability and ease of bioinformatics protocol verification, a standard reporting format (for example, a BioCompute Object (BCO)) should be used to record all thresholds, steps and implementation arguments for a given pipeline[58]. Comparisons of BCOs from different pipelines can then be used to set acceptable lower limits for the assessed parameters, refining technical validation criteria across pipelines[59].

A prime example of external quality control of bioinformatics pipelines is the efforts by the Netherlands National Institute for Public Health and the Environment (RIVM) to standardize the use of WGS for MTBC genotyping across the European Reference Laboratory Network for TB (ERLTB-Net). Panels of DNA extracted

WGS workflows
All steps involved (from culturing to SNP calling and analyses) for whole genome sequencing of an isolate.

BioCompute Object
(BCO). A framework for standardized reporting of computational parameters for a whole genome sequencing pipeline.

from selected MTBC isolates are sent annually by the RIVM to reference laboratories to assess intralaboratory and interlaboratory reproducibility of WGS. Similar efforts in high-burden settings are needed to monitor the reliability of MTBC WGS outputs when used in these settings.

Validation for core tasks: transmission, phylogeny and drug resistance. Task validation is used to demonstrate that a given pipeline is verified for a specific analysis (for example, drug resistance profiling). For task validation, MTBC bioinformatics pipelines should use defined validation data sets, ideally with hundreds or thousands of well-characterized clinical MTBC strains representing the diversity of a specific core task (for example, different drug susceptibility profiles for resistance detection, representatives of all MTBC phylogenetic diversity for typing or differing degrees of clustering for transmission analyses). The number of readily available, well-curated validation data sets is currently limited.

For validation of transmission cluster detection, the RIVM has provided laboratories with sequenced reads from 535 MTBC isolates for which epidemiological links are known. Using this data set, the EUSeqMyTB consortium showed that existing pipelines could confidently distinguish linked from unlinked cases, especially when the SNP distances are high, as is often the case in low-burden settings[12]. This comparison was undertaken as part of an effort to standardize WGS for monitoring cross-border transmission of multidrug-resistant TB in Europe[15].

The clonality of MTBC strains means that lineage and strain typing can be performed using only a handful of SNPs that are specific for strains of a particular lineage. Several studies have demonstrated the reliability of specific SNPs to determine the MTBC lineage or sublineage[8,9,60]. However, sublineage classifications are often less resolved, and parallel nomenclatures for lineage 2 are being used[18,61,62]. As the diversity of the MTBC is further explored, especially for animal-associated and zoonotic TB, these underdescribed lineages can also easily be strain typed using the same SNP-based approach[7].

Validation of WGS for drug resistance is the most advanced of all the core tasks. Studies showed high concordance between phenotypic and genotypic predictions, regardless of the sequencing platform used[19,53]. In the past 2 years, major progress has been made in the linkage between genotype and resistance phenotype by use of a standardized statistical approach[24,25]. The task of incrementally improving our knowledge base on genetic resistance profiling is primarily being addressed by the two global consortia outlined earlier: ReSeqTB's single platform for genotype–phenotype investigation of drug resistance[26,34] and CRyPTIC's genotypic–phenotypic linking of more than 10,000 isolates demonstrating susceptibility prediction with 99% sensitivity for rifampicin and isoniazid and 93–96% for ethambutol and pyrazinamide[2]. These results have led to some low-burden countries (for example, the Netherlands and the United Kingdom) replacing phenotypic DST with WGS-based DST for first-line drugs. Resistance predictions for drugs used to treat multidrug-resistant TB can also be undertaken with sensitivity often ~90%[24]. Large comparative studies using phenotype–genotype associations are expanding the catalogue of resistance markers[63,64] and will help to increase the sensitivity for detecting multidrug-resistant TB. Efforts are now directed towards increasing the diversity of isolates and including accompanying high-quality phenotypic and clinical data, especially for new anti-TB drugs.

Standardization of communication of MTBC WGS results and data sharing. Effective communication of WGS-based results to a diverse audience of end users is key to positively impacting patient care and TB control programmes. Although the need for plain language reporting of genomic results has been recognized[51,65], there are no international standards yet. Reporting standards should be flexible enough to address the differing levels of familiarity of end users with genomic data interpretation and allow customization to region-specific treatment guidelines and formatting requirements. For example, the International Organization for Standardization ISO 15189:2012 standard mandates that information such as patient identifiers, assay details and the testing laboratory is reported. Recommendations from MTBC WGS report design validation studies include avoiding the use of abbreviations, drawing attention to important elements with shading, bolding and other types of emphasis, and incorporating summary statements to effectively communicate key results[66,67].

In peer-reviewed publications, the parameters used at each step of a bioinformatics pipeline must be stated in a manner that makes the analysis reproducible and understandable to non-bioinformaticians (for example, using a BCO as outlined earlier). Custom code used in the analysis should be made available through a public repository (for example, GitHub), ensuring ease of installation elsewhere. Pipelines should report the outcome of technical validations, at least for the core tasks that they aim to address (for example, lineage-defining SNPs for a strain typing pipeline). Examples of standard reporting include the Minimum Information About a Bioinformatics Investigation[68] and the Strengthening the Reporting of Molecular Epidemiology for Infectious Diseases[69] guidelines. In Supplementary Box 1, we suggest data elements to include according to intended use, but note that a report may need to include elements from more than one use case.

Data sharing is crucial as incremental knowledge improves drug resistance predictions and strain tracking relies on the number and diversity of strain genome data available. Data that are shared include either coded strain identifiers such as MLST patterns or raw sequence data not yet processed by a pipeline. Data sharing has already been shown to be invaluable for detecting cross-Europe transmission clusters[14]. Data sharing should encompass data produced by research and collected in public health laboratories and from surveillance efforts[70], similarly to what is done by the GenomeTrakr network for foodborne pathogens[71], while safeguarding patient data and appropriately acknowledging contributions.

Spoligotyping
A PCR-based approach based on the amplification of spacers in the CRISPR region of the *Mycobacterium tuberculosis* complex (MTBC). It is used for genotyping MTBC strains.

The crucial next step for fully utilizing MTBC WGS data is implementation of validations, both technical and task oriented, for all pipelines. Once undertaken, the agreed-upon pipeline or pipelines can then be widely implemented, once infrastructure and usability are accounted for.

Implementation of WGS

Although WGS is becoming widely used in research, minimal progress has been made in the implementation of WGS in clinical and public health applications. Some reasons include the lack of standardized end-to-end solutions, the required wet-laboratory and computing infrastructure, the need for sufficient Internet connectivity and bandwidth, and training deficits in genomics and bioinformatics[72–74]. Efforts are thus needed to expand accessibility to perform analysis by non-experts. How these factors are addressed will depend on a country's income and public health sector strength.

High-income countries will probably use a mixture of closed (end-to-end) solutions and more complex pipelines as they likely will have on-site bioinformatics support. Ideally, routine analysis of WGS data will require little to no bioinformatics knowledge by the end user. Implementation of these pipelines can be undertaken by either local set-ups with supporting infrastructure or a Web-based approach with easy, affordable access[75]. Many large healthcare facilities, such as referral hospitals, are already incorporating bioinformatics units into their support services as part of the trend towards personalized medicine, a practice that TB treatment can take advantage of. These services should mediate the implementation of complex pipelines and make all required software readily available without a requirement to install additional software tools, as is done with certain existing pipelines[33,47].

Given the heterogeneity of pipelines already in place (Supplementary Table 1), it is conceivable that pipelines will become more numerous and diverse when implementation is done in hundreds of care services. Some will opt in for end-to-end solutions, perhaps integrated with the sequencing platform, whereas others will prefer task-specific pipelines, such as resistance prediction only. Those implementing their own pipeline should be aware of the limitations, cautions and recommendations detailed by expert consensus here and elsewhere[6,75]. To evaluate new pipelines it is preferable to develop inside 'containers' (that is, cross-platform, stand-alone software sections that contain the pipeline and all required dependencies), such as Docker or Singularity[76,77], or package managers such as Bioconda or Homebrew which allow easy installation of platform-specific programs[78,79]. Creating a container for each step (FIG. 2) also allows easy updating of a specific step without the need to install a whole new pipeline and allows tasks (for example, resistance profiling) to be added to the pipeline as needed. To allow usability by a range of end users, high-level access to the individual steps should be available for advanced users, with functionality layers abstracted away from users with limited bioinformatics expertise. The pipelines should be open source and user-friendly, by using intuitive and well-documented command line and graphical user interfaces with relevant and validated default parameters.

The situation in low- and middle-income (LMIC) countries, especially those with a high burden of TB, is currently totally different. End-to-end solutions based on cloud computing are the most logical step forward, similar to the roll-out of quantitative PCR systems (BOX 1). Centralized Web-based analysis platforms have recently emerged and promise to aid in computational efficiency, access and usability[46,50]. Roll-out of such initiatives to more countries would greatly increase the potential for large-scale WGS implementation. The primary barrier to this is usually unstable Internet connectivity with limited bandwidth, although use of methods that can effectively handle connection interruptions, such as BioTorrents[80], or direct transfer from sequencing centres to cloud storage and/or Web-based pipelines may help circumvent these issues.

The use of end-to-end, cloud-based solutions is likely to have an important role in LMIC countries. It is, however, advisable to train more scientists in MTBC WGS in those countries[81,82]. Although standardized, unchangeable pipelines are optimal for global implementation of WGS, there are several reasons why local bioinformatics knowledge is required, such as the necessity to adapt analyses to the country-specific epidemiological profiles and public health ecosystems or regulatory laws that do not allow storage beyond country borders. Such customized, yet reproducible solutions are being

Box 1 | **Primary *Mycobacterium tuberculosis* diagnostics**

Solid or liquid culture (for example, Mycobacteria Growth Indicator Tube[137]) is the conventional diagnostic tool for *Mycobacterium tuberculosis* complex (MTBC) identification and drug susceptibility testing. However, such phenotypic tests can take weeks to months to obtain results, require high-level biosafety infrastructure and are considered unreliable for certain drugs (for example, pyrazinamide). Therefore, several molecular tests (besides whole genome sequencing) that are directly applicable to clinical samples have been developed. Line probe assays rely on the hybridization of amplified mycobacterial DNA to nucleotide probes on strips to detect select drug resistance-associated mutations or their wild type alleles. The line probe assays MTBDRplus[138,139], TB NTM+MDR[139,140] and MTBDRsl[141,142] were all endorsed by the WHO. The two former assays identify mutations associated with resistance to rifampicin (in *rpoB*) and isoniazid (in *katG* and *inhA*); that is, they detect multidrug-resistant tuberculosis. The MTBDRsl assay identifies mutations associated with resistance to fluoroquinolones (in *gyrA* and *gyrB*) and aminoglycosides (in *rrs* and *eis*); that is, it detects extensively drug resistant tuberculosis. Other tests use (cartridge-based) real-time PCR (GeneXpert MTB/RIF[86,143] (and updated Ultra[144,145]), Anyplex II MTB/MDR/XDR[146] and FluoroType MTBDR[147]) or PCR melting curve (Meltpro[148]) for mutation detection. FluoroType and the WHO-endorsed and globally deployed GeneXpert both detect rifampicin-associated mutations in *rpoB* and, in the case of FluoroType, isoniazid resistance mutations (in *katG*, and *inhA*). As all aforementioned molecular tests use indirect sequencing technologies, they are intrinsically limited to the detection of common preselected mutations and are prone to false positive results due to indiscriminate detection of unrelated mutations[149,150]. To circumvent these limitations, newer assays use targeted amplicon sequencing. The Next Gen-RDST[151,152] and Deeplex-MycTB[153,154] assays are directly applicable to clinical samples and sequence 6 or 18 genes (including some promoter regions) associated with resistance to 7 or 13 antituberculosis drugs, respectively. Deeplex-MycTB additionally includes identification of mycobacterial species and uses spoligotyping (spacer oligonucleotide typing). The large read coverage depths that can be achieved with Deeplex-MycTB allow high-confidence mutation calls, including those born by minor subpopulations in the case of heteroresistance. Nevertheless, the accessible targets are inherently fewer than with whole genome sequencing.

supported by capacity-building initiatives (for example, the Human, Heredity and Health in Africa Consortium and the TORCH consortium). TB supranational reference laboratories should also have an important coordinating role, as is currently done for phenotypic workflows[19,83]. Ultimately, expansion of education curricula to include bioinformatics is needed to generate sufficient capacity[84].

Finally, supportive policy and political commitment will be essential for sustainable implementation of WGS, especially in TB-endemic LMIC countries[73,81,85]. This implementation will benefit from the lessons learned during the stepwise approach used to roll out line probe assays and GeneXpert[86] (BOX 1).

Extensions of the current standard

Although current pipelines (FIG. 2) appear to be highly accurate for many aspects of the three core tasks, multiple important issues remain open and should be part of future research and evaluation.

Input data validation and quality control. Most current pipelines do not routinely filter out reads that do not come from MTBC strains. However, sequencing files can contain reads from other organisms, and these contaminants can introduce errors during the variant calling process, modifying both the variants identified and their respective frequencies[87]. Additionally, any host DNA sequencing reads should be removed for legal or ethical reasons, especially if the data are shared online. Computationally removing non-MTBC strain reads before mapping is an efficient strategy to implement contamination-proof analysis pipelines[39] but requires taxonomic classification of individual reads. Use of taxonomic classification methods, in which reads are assigned to the closest matched species, allows quick and efficient removal of contaminating reads but requires comprehensive genome databases, often making their implementation extremely memory consuming[88,89]. Additionally, elimination of reads from highly conserved core bacterial genes of heterologous sources remains a problem. Proposed alternatives include masking genomic regions known to accumulate artefactual polymorphisms[87], filtering the alignments produced by contaminant reads, or fine-tuning the read aligners such that only the MTBC strains sequences are mapped to the reference genome. Any method will require thorough technical validation to ensure that contaminant reads are removed without eliminating true MTBC sequences (for example, through in silico generation of data sets with differing levels of reads from other organisms).

Sequence read mapping and reference genomes. The use of a single reference genome for mapping all MTBC strains is the ideal approach for comparable and standard variant calling. Although most pipelines use the *M. tuberculosis* strain H37Rv genome[4,90] as the reference genome, several alternative approaches should be explored. As H37Rv is a lineage 4 strain, its use as a reference for other lineages may be inappropriate due to gene content differences between lineages[91–94]. Additionally, H37Rv contains many variants not found in any other strain[95], including in genes related to drug resistance (for example, *gyrA* S95T), creating confusion in SNP interpretations. Any replacement of H37Rv as the reference genome should be assessed by in silico studies across data sets and clinical settings. An example of such a study tested seven different references against sequence reads from lineage 4 isolates and showed that very limited variation occurred and that reference choice should be based on criteria other than matching lineage[96].

One alternative to the H37Rv genome is a pangenome which incorporates the entire gene pool of MTBC lineages. Studies have found small but notable differences in gene content between lineages, often affecting genes involved in pathogenesis[91–94]. Although these differences are unlikely to affect drug resistance profiling (because associated mutations are in the core genome), they may impact delineation of transmission clusters if additional SNPs are found in these genes that would push strain comparisons over the predetermined thresholds. Building an MTBC pangenome should be straightforward due to the close genetic relationship between different strains (average nucleotide identity between any two strains ≥99.8%) and the lack of horizontal gene transfers events. So far this approach has not been effectively explored.

A second alternative is the use of an inferred ancestral genome representative of the MTBC population and diversity[28,39]. From an evolutionary perspective, this approach addresses the *M. tuberculosis* strain H37Rv-specific variants outlined above. In addition, because all extant strains are equidistant to a common ancestor, the number of SNPs called for any MTBC strain will be similar (normalized) regardless of its lineage. This expected SNP range is useful for quality control, as deviations may indicate poor quality sequencing, mixed infections or contaminations[39].

A third approach is to use ad hoc reference genomes, depending on the study being conducted. For instance, lineage-specific ancestral genomes or high-quality, closed, outbreak-specific reference genomes[97–99] could be used as a reference to reduce mapping errors[10]. A disadvantage of this approach is that it hampers comparison of results between pipelines and the standardized reporting of results.

A completely different alternative involves de novo assembly, using a reference-free approach, which has been successfully applied for human population genomics data[100].

Independent of the selection of the reference genome, other steps such as mapping and filtering are not consistent between different pipelines but might greatly affect the analysis outcome. For instance, removal of duplicate reads may have a large impact on the variants identified and allele frequencies. Similarly, local assembly or realignment around indels, reducing false positive SNPs derived from mapping artefacts, is rarely used in MTBC WGS pipelines[57] but is known to affect variant calling[46]. The question of whether these steps have a relevant effect on the final outcome should be incorporated into future technical validations.

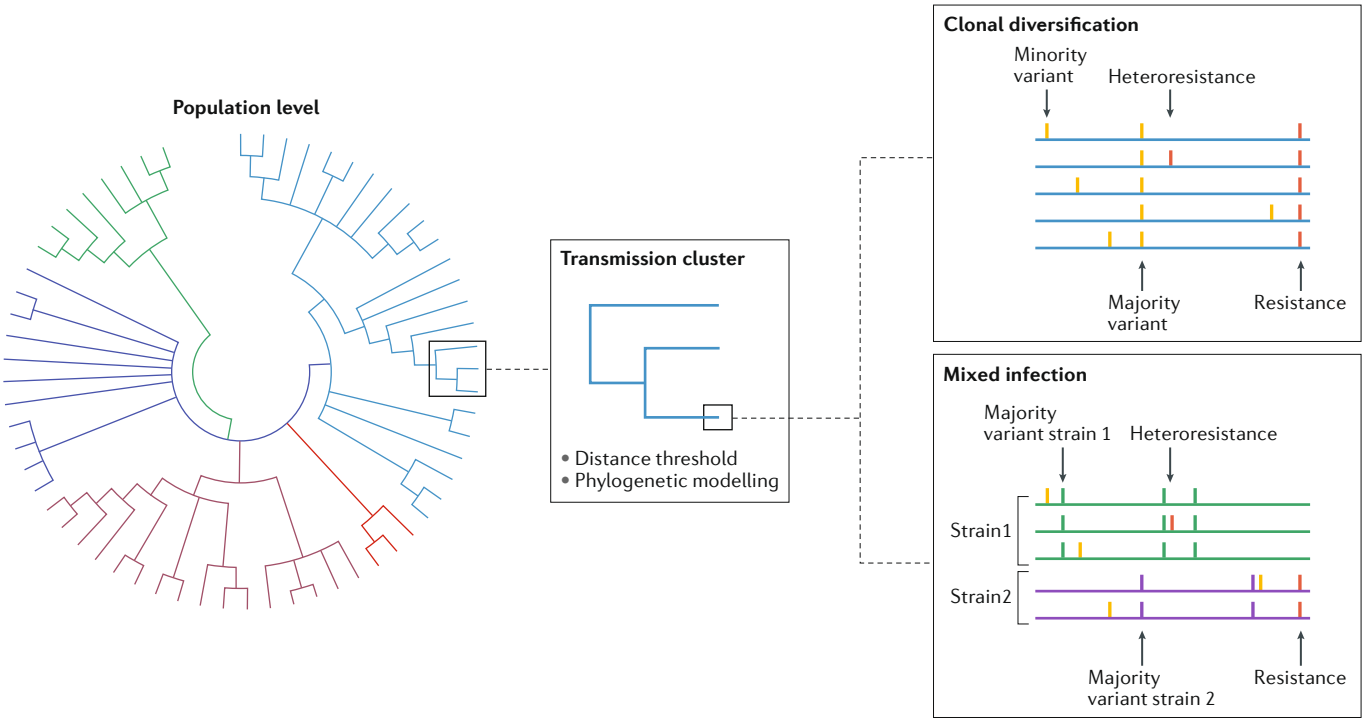


Fig. 4 | **Epidemiological and within-host applications of SNP-based comparisons between *Mycobacterium tuberculosis* complex isolates.** Whole genome sequencing data can be used at multiple levels of epidemiological complexity. At a global population level, SNP-based phylogenetics can be used to delineate strains and subspecies within the *M. tuberculosis* complex. At the local level, these phylogenies can be subdivided into transmission clusters using predefined SNP or allele cut-offs. Finally, at the individual level, within-host diversity can be examined using SNP proportions to detect heteroresistance (subpopulations with different drug resistance profiles) or mixed infections (a single host infected multiple times).

Interpretation of drug resistance results and predictions. Currently, the bulk of routine drug resistance testing is undertaken using phenotypic DST. This approach will still be required for a subset of difficult-to-interpret drug resistance patterns; however, the overarching goal is to detect all variants associated with resistance for comprehensive genome-based resistance profiling. Although the current statistical approach for identifying resistance-associated variants using WGS data is an important step forward for clinical use, a weakness is that phenotype predictions of rare and/or novel genetic variants cannot be assessed (FIG. 3). This problem is especially relevant for identifying resistance to new and repurposed drugs, or drugs such as pyrazinamide and ethionamide for which drug resistance mutations do not arise in hotspots but appear across entire genes (for example, *pncA* and *ethA*) and in promoter regions. For uncommon or novel genomic variants, the standard statistical approach could be complemented by experimental data, comprehensive single-nucleotide mutagenesis[101] followed by systematic phenotypic screening, multiomics studies and machine learning approaches to predict the resistance phenotype[102,103]. With the final aim of replacing most of phenotypic DST with sequence-based testing, it will also be essential to catalogue 'benign' variants that are not associated with resistance (that is, phylogenetic markers or other neutral variants[2]). New statistical approaches such as large-scale genome-wide association studies[63,64], protein structure modelling[43,104] and machine learning[102,103,105] will likely have a key role in identifying causative versus benign variants. Comprehensive databases of WGS data linked with phenotypic and clinical outcome data (for example, CRyPTIC or ReSeqTB) are key to moving towards this goal.

Once established, endorsement of a single standardized variant list by the WHO or other regulatory body with regular updating should be pursued.

Variant calling for other purposes. Accurate variant calling has major implications on downstream interpretation of the results for evolutionary, epidemiological and clinical applications. Owing to the low levels of diversity and the slow substitution rate of MTBC genomes[31,41,98,106], a few falsely called SNPs can affect the interpretation of transmission events, lead to the false diagnosis of a relapsed infection as reinfection or influence the interpretation of subpopulations within a patient (FIG. 4).

A primary use of MTBC WGS is the identification of recent transmission chains and their direction at high resolution. Although some studies have used thresholds ranging from 0 to 50 SNPs[107–109], a threshold of 5 or 12 SNPs is most frequently used to identify possible epidemiological links and recent transmission[29,31]. For WGS-based distinction of relapse versus reinfection, studies have often used arbitrary thresholds of less than

6 or less than 10 SNPs to define reactivation, and more than 100 SNPs to define reinfection[45,110,111]. Any threshold selection can be problematic as inferences based on relatedness must include possible underlying methodological bias (culture, sampling and pipeline). In addition, genetic distances may be impacted by biological factors such as potential mutational bursts[41,112], clonal variants in different lesions[10,113], the impact of strain type (lineage or subspecies) or drug resistance on substitution rates[106,114], and genome instability during latency[114,115]. For example, identifying transmission from unrelated cases or distinguishing relapse and reinfection in low-burden countries is relatively easy, where the distribution of SNP distances is bimodal, separating linked from unlinked cases[12,14]. Conversely, inferring transmission clusters within the context of institutional or household settings or in high-TB-incidence scenarios where the SNP distance distribution is continuous remains difficult, especially if epidemiological links in large clusters of patients with seemingly identical strains are lacking[116–118].

Other approaches beyond SNP-based clustering have been developed to improve the identification of epidemiological links and improve the resolution of transmission networks in outbreaks. These either use transmission event thresholds[119] or combine genomic and epidemiological data to identify the most probable transmission trees for infectious diseases, or do both[120,121]. One particularly important consideration when one is reconstructing transmission networks of MTBC outbreaks is that phylogeny and transmission events do not necessarily coincide as a consequence of genetic diversification during latency and long generation times[122]; it is thus necessary to model the within-host genetic dynamics[123–125]. Besides transmission reconstruction, phylodynamic approaches also allow the inference of epidemiologically relevant parameters such as the effective reproduction number, as well as the timing and geographic origin of an outbreak[126,127].

Within-host diversity and subpopulation detection remains even more challenging. Low-frequency variants that are not due to technical artefacts can indicate the presence of mixed infections (that is, coinfection with two distinct MTBC strains), or microevolution leading to closely related subpopulations, or heteroresistance (subpopulations that differ in drug resistance-related variants)[10,113,128]. Proposed subpopulation detection limits in different pipelines range from 10% to 75% (Supplementary Table 1) and are strongly influenced by factors such as read depth. Although the presence of a subpopulation of at least 1% resistant bacilli is considered clinically relevant[129], selection bias means that what is observed in sequencing data may not be representative of what is present in the culture isolate, which in turn is likely not representative of the diversity in the sputum sample, which is known to not represent the entirety of the within-patient diversity[113,130]. Mathematical modelling approaches have been developed to identify mixed infections[131,132]. However, with current approaches the detection of mixed infections is limited by the relative ratio of the two strains and the number of differing SNPs. Further research and methodological improvements are needed to better understand and interpret this within-host diversity.

Effective reproduction number
The average number of secondary cases per infectious case.

Microevolution
Genetic changes within a population, resulting in separate subpopulations.

Beyond the current standards

As current culture-based approaches require time for MTBC strain growth, culture-free WGS directly from clinical samples (for example, sputum) would be transformative for clinical and public health applications of WGS. This approach would not only eliminate the culture delay but also remove culture selection biases. Although studies have shown some success, this approach is still mired with problems such as contamination by human and commensal microbial reads, preventing sufficient coverage depth of the MTBC genomes and thus reliable variant calling, even in samples with high bacterial loads[133–135]. Improvements in cell lysis or capture coupled with selective DNA enrichment or depletion could reduce this technical complexity and cost. Additionally, downstream bioinformatic filtering could be used to control for and remove possible remaining false variants.

Much is expected from the development of highly portable sequencing devices (for example, the MinION). Such technology offers the capacity to detect variants in real time during sample acquisition, potentially reporting results from sputum within hours if mycobacterial loads are high. Their portability and ability to work in resource limited settings also favours direct sequencing of clinical samples, even in LMIC countries. Moreover, although progress has been made in analysis of variants in repeat-rich genome regions (for example, PE and PPE gene families) or structural changes (duplications, large indels and so on) by short read mapping[110,136], long read sequencing will make this more robust[99,133]. Unfortunately, application of this technology is currently limited by high error rates (although new dual sequence reading systems promise substantial improvement) and, specifically for mycobacteria, difficulty in cell lysis without overshearing DNA.

Conclusions

A decade after the first proof-of-principle studies, the community consensus is that MTBC WGS is now advanced enough to inform clinical decisions and public health. This is evident as WGS has already replaced phenotypic testing for first-line drugs in some settings, has become the basis of drug resistance surveillance surveys supported by the WHO and has become the standard for MTBC molecular epidemiology and strain typing studies. Before its full-scale implementation, we call for extensive standardization and validation efforts. This will require political commitment and the involvement of supranational laboratories and regulatory authorities. There also remains an important role for the research community at large to continue to improve the technical and analytical aspects of WGS. Consideration is also needed for the ethical implications and consequences of routine WGS and the information it provides. There is therefore a need now to commit resources to ensure access to standardized and validated WGS approaches, especially in high-burden countries, where WGS will have the greatest impact.

Published online: 17 June 2019

1. World Health Organization. Global tuberculosis report 2018. *WHO* https://www.who.int/tb/publications/global_report/archive/ (2018).
2. The CRyPTIC Consortium and the 100,000 Genomes Project. Prediction of susceptibility to first-line tuberculosis drugs by DNA sequencing. *N. Engl. J. Med.* **379**, 1403–1415 (2018).
The first large-scale study demonstrating how phenotypic testing can be replaced by WGS for first-line drug testing.
3. Gardy, J. L. et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N. Engl. J. Med.* **364**, 730–739 (2011).
4. Cole, S. T. et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544 (1998).
5. Cabibbe, A. M., Walker, T. M., Niemann, S. & Cirillo, D. M. Whole genome sequencing of *Mycobacterium tuberculosis*. *Eur. Respir. J.* **52**, 1801163 (2018).
6. Satta, G. et al. *Mycobacterium tuberculosis* and whole-genome sequencing: how close are we to unleashing its full potential? *Clin. Microbiol. Infect.* **24**, 604–609 (2018).
An extensive review of the literature outlining the potential of WGS for TB research and clinical use.
7. Lipworth, S. et al. SNP-IT tool for identifying subspecies and associated lineages of *Mycobacterium tuberculosis* complex. *Emerg. Infect. Dis.* **25**, 482–488 (2019).
8. Coll, F. et al. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat. Commun.* **5**, 4812 (2014).
This study reports the now standard sublineage typing scheme using SNP-based information for MTBC.
9. Homolka, S. et al. High resolution discrimination of clinical *Mycobacterium tuberculosis* complex strains based on single nucleotide polymorphisms. *PLOS ONE* **7**, e39855 (2012).
10. Trauner, A. et al. The within-host population dynamics of *Mycobacterium tuberculosis* vary with treatment efficacy. *Genome Biol.* **18**, 71 (2017).
11. Merker, M., Kohl, T. A., Niemann, S. & Supply, P. The evolution of strain typing in the *Mycobacterium tuberculosis* complex. *Adv. Exp. Med. Biol.* **1019**, 43–78 (2017).
12. Jajou, R. et al. Epidemiological links between tuberculosis cases identified twice as efficiently by whole genome sequencing than conventional molecular typing: A population-based study. *PLOS ONE* **13**, e0195413 (2018).
This study shows the advantage of WGS approaches over mycobacterial interspersed repetitive unit variable-number tandem repeat genotyping for detection of transmission clusters.
13. Wyllie, D. H. et al. A quantitative evaluation of MIRU-VNTR typing against whole-genome sequencing for identifying *Mycobacterium tuberculosis* transmission: a prospective observational cohort study. *EBioMedicine* **34**, 122–130 (2018).
14. Walker, T. M. et al. A cluster of multidrug-resistant *Mycobacterium tuberculosis* among patients arriving in Europe from the Horn of Africa: a molecular epidemiological study. *Lancet Infect. Dis.* **18**, 431–440 (2018).
15. Tagliani, E. et al. EUSeqMyTB to set standards and build capacity for whole genome sequencing for tuberculosis in the EU. *Lancet Infect. Dis.* **18**, 377 (2018).
Announcement of the European Centre for Disease Prevention and Control efforts to establish and validate the use of WGS for all TB public health initiatives.
16. Cohen, K. A. et al. Evolution of extensively drug-resistant tuberculosis over four decades: whole genome sequencing and dating analysis of *Mycobacterium tuberculosis* isolates from KwaZulu-Natal. *PLOS Med.* **12**, e1001880 (2015).
17. Eldholm, V. et al. Four decades of transmission of a multidrug-resistant *Mycobacterium tuberculosis* outbreak strain. *Nat. Commun.* **6**, 7119 (2015).
18. Merker, M. et al. Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. *Nat. Genet.* **47**, 242–249 (2015).
19. Zignol, M. 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.* **18**, 675–683 (2018).
20. Gröschel, M. I. et al. Pathogen-based precision medicine for drug-resistant tuberculosis. *PLOS Pathog.* **14**, e1007297 (2018).
21. World Health Organization. The use of next-generation sequencing technologies for the detection of mutations associated with drug resistance in *Mycobacterium tuberculosis* complex: technical guide. *WHO* https://apps.who.int/iris/handle/10665/274443 (2018).
This guide is the first step towards validation of WGS as a tool for MTBC clinical and public health work.
22. Nebenzahl-Guimaraes, H., Jacobson, K. R., Farhat, M. R. & Murray, M. B. Systematic review of allelic exchange experiments aimed at identifying mutations that confer drug resistance in *Mycobacterium tuberculosis*. *J. Antimicrob. Chemother.* **69**, 331–342 (2014).
23. Sandgren, A. et al. Tuberculosis drug resistance mutation database. *PLOS Med.* **6**, e1000002 (2009).
24. Coll, F. et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med.* **7**, 51 (2015).
25. Miotto, P. et al. A standardised method for interpreting the association between mutations and phenotypic drug resistance in *Mycobacterium tuberculosis*. *Eur. Respir. J.* **50**, 1701354 (2017).
This was the first study to use a score system to classify mutations for clinical interpretation.
26. Starks, A. M. et al. Collaborative effort for a centralized worldwide tuberculosis relational sequencing data platform. *Clin. Infect. Dis.* **61**, S141–S146 (2015).
This publication outlines the design and use of the ReSeqTB platform.
27. Brown, T., Nikolayevskyy, V., Velji, P. & Drobniewski, F. Associations between *Mycobacterium tuberculosis* strains and phenotypes. *Emerg. Infect. Dis.* **16**, 272–280 (2010).
28. Comas, I. et al. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nat. Genet.* **45**, 1176–1182 (2013).
29. Meehan, C. J. et al. The relationship between transmission time and clustering methods in *Mycobacterium tuberculosis* epidemiology. *EBioMedicine* **37**, 410–416 (2018).
30. Kohl, T. A. et al. Harmonized genome wide typing of tubercle bacilli using a web-based gene-by-gene nomenclature system. *EBioMedicine* **34**, 131–138 (2018).
31. Walker, T. M. et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect. Dis.* **13**, 137–146 (2013).
32. Koster, K. J. et al. Genomic sequencing is required for identification of tuberculosis transmission in Hawaii. *BMC Infect. Dis.* **18**, 608 (2018).
33. Kohl, T. A. et al. MTBseq: a comprehensive pipeline for whole genome sequence analysis of *Mycobacterium tuberculosis* complex isolates. *PeerJ* **6**, e5895 (2018).
34. Ezewudo, M. et al. Integrating standardized whole genome sequence analysis with a global *Mycobacterium tuberculosis* antibiotic resistance knowledgebase. *Sci. Rep.* **8**, 15382 (2018).
35. Brynildsrud, O. B. et al. Global expansion of *Mycobacterium tuberculosis* lineage 4 shaped by colonial migration and local adaptation. *Sci. Adv.* **4**, eaat5869 (2018).
36. Brown, A. C. et al. Rapid whole-genome sequencing of *Mycobacterium tuberculosis* isolates directly from clinical samples. *J. Clin. Microbiol.* **53**, 2230–2237 (2015).
37. Conceição, E. C. et al. Analysis of potential household transmission events of tuberculosis in the city of Belem, Brazil. *Tuberculosis* **113**, 125–129 (2018).
38. Walker, T. M. et al. Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect. Dis.* **15**, 1193–1202 (2015).
39. Goig, G. A., Blanco, S., Garcia-Basteiro, A. & Comas, I. Pervasive contaminations in sequencing experiments are a major source of false genetic variability: a *Mycobacterium tuberculosis* meta-analysis. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/403824v1 (2018).
40. Menardo, F. et al. Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics* **19**, 164 (2018).
41. Bryant, J. M. et al. Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. *BMC Infect. Dis.* **13**, 110 (2013).
42. Shea, J. 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.* **55**, 1871–1882 (2017).
43. Phelan, J. et al. *Mycobacterium tuberculosis* whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC Med.* **14**, 31 (2016).
44. Witney, A. A. et al. Use of whole-genome sequencing to distinguish relapse from reinfection in a completed tuberculosis clinical trial. *BMC Med.* **15**, 71 (2017).
45. Casali, N. et al. Whole genome sequence analysis of a large isoniazid-resistant tuberculosis outbreak in London: a retrospective observational study. *PLOS Med.* **13**, e1002137 (2016).
46. Feuerriegel, S. et al. PhyResSE: a web tool delineating *Mycobacterium tuberculosis* antibiotic resistance and lineage from whole-genome sequencing data. *J. Clin. Microbiol.* **53**, 1908–1914 (2015).
47. Bradley, P. et al. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. *Nat. Commun.* **6**, 10063 (2015).
48. Iwai, H., Kato-Miyazawa, M., Kirikae, T. & Miyoshi-Akiyama, T. CASTB (the comprehensive analysis server for the *Mycobacterium tuberculosis* complex): a publicly accessible web server for epidemiological analyses, drug-resistance prediction and phylogenetic comparison of clinical isolates. *Tuberculosis* **95**, 843–844 (2015).
49. Steiner, A., Stucki, D., Coscolla, M., Borrell, S. & Gagneux, S. KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. *BMC Genomics* **15**, 881 (2014).
50. Farhat, M. et al. genTB: translational genomics of tuberculosis. *genTB* https://gentb.hms.harvard.edu (2015).
51. Schleusener, V., Köser, C. U., Beckert, P., Niemann, S. & Feuerriegel, S. *Mycobacterium tuberculosis* resistance prediction and lineage classification from genome sequencing: comparison of automated analysis tools. *Sci. Rep.* **7**, 46327 (2017).
52. Ngo, T.-M. & Teo, Y.-Y. Genomic prediction of tuberculosis drug-resistance: benchmarking existing databases and prediction algorithms. *BMC Bioinformatics* **20**, 68 (2019).
53. Phelan, J. et al. The variability and reproducibility of whole genome sequencing technology for detecting resistance to anti-tuberculous drugs. *Genome Med.* **8**, 132 (2016).
54. Macedo, R. et al. Dissecting whole-genome sequencing-based online tools for predicting resistance in *Mycobacterium tuberculosis*: can we use them for clinical decision guidance? *Tuberculosis* **110**, 44–51 (2018).
55. Angers-Loustau, A. et al. The challenges of designing a benchmark strategy for bioinformatics pipelines in the identification of antimicrobial resistance determinants using next generation sequencing technologies. *F1000Res.* **7**, 459 (2018).
56. US Food and Drug Administration. Infectious disease next generation sequencing based diagnostic devices: microbial identification and detection of antimicrobial resistance and virulence markers. *FederalRegister.gov* https://www.federalregister.gov/documents/2016/08/11/2016-19109/infectious-disease-next-generation-sequencing-based-diagnostic-devices-microbial-identification-and (2016).
57. Pouseele, H. & Supply, P. Accurate whole-genome sequencing-based epidemiological surveillance of *Mycobacterium tuberculosis*. *Methods Microbiol.* **42**, 359–394 (2015).
58. Simonyan, V., Goecks, J. & Mazumder, R. Biocompute objects — a step towards evaluation and validation of biomedical scientific computations. *PDA J. Pharm. Sci. Technol.* **71**, 136–146 (2017).
59. Alterovitz, G. et al. Enabling precision medicine via standard communication of HTS provenance, analysis, and results. *PLOS Biol.* **16**, e3000099 (2018).
60. Stucki, D. et al. Standard genotyping overestimates transmission of *Mycobacterium tuberculosis* among immigrants in a low-incidence country. *J. Clin. Microbiol.* **54**, 1862–1870 (2016).
61. Liu, Q. et al. China's tuberculosis epidemic stems from historical expansion of four strains of *Mycobacterium tuberculosis*. *Nat. Ecol. Evol.* **2**, 1982–1992 (2018).
62. Holt, K. E. et al. Frequent transmission of the *Mycobacterium tuberculosis* Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. *Nat. Genet.* **50**, 849–856 (2018).
63. Coll, F. et al. Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis*. *Nat. Genet.* **50**, 307–316 (2018).
64. Farhat, M. R. et al. GWAS for quantitative resistance phenotypes in *Mycobacterium tuberculosis* reveals resistance genes and regulatory regions.

Nat. Commun. https://doi.org/10.1038/s41467-019-10110-6 (2019).
65. Kwong, J. C., Mccallum, N., Sintchenko, V. & Howden, B. P. Whole genome sequencing in clinical and public health microbiology. *Pathology* **47**, 199–210 (2015).
66. Crisan, A., McKee, G., Munzner, T. & Gardy, J. L. Evidence-based design and evaluation of a whole genome sequencing clinical report for the reference microbiology laboratory. *PeerJ* **6**, e4218 (2017). **This article reports new standards for reporting of WGS-based TB clinical information**.
67. Tornheim, J. A. et al. Building the framework for standardized clinical laboratory reporting of next generation sequencing data for resistance-associated mutations in *Mycobacterium tuberculosis* complex. Clin. Infect. Dis. https://doi.org/10.1093/cid/ciz219 (2019).
68. Tan, T. W. et al. Advancing standards for bioinformatics activities: persistence, reproducibility, disambiguation and Minimum Information About a Bioinformatics investigation (MIABi). *BMC Genomics* **11** (Suppl. 4), 27 (2010).
69. Field, N. et al. Strengthening the Reporting of Molecular Epidemiology for Infectious Diseases (STROME-ID): an extension of the STROBE statement. *Lancet Infect. Dis.* **14**, 341–352 (2014).
70. World Health Organization. WHO's code of conduct for open and timely sharing of pathogen genetic sequence data during outbreaks of infectious disease. *WHO* https://www.who.int/blueprint/what/norms-standards/gsdsharing/en/ (2019).
71. Allard, M. W. et al. Practical value of food pathogen traceability through building a whole-genome sequencing network and database. *J. Clin. Microbiol.* **54**, 1975–1983 (2016).
72. Karikari, T. K. Bioinformatics in Africa: the rise of Ghana? *PLOS Comput. Biol.* **11**, e1004308 (2015).
73. Tekola-Ayele, F. & Rotimi, C. N. Translational genomics in low- and middle-income countries: opportunities and challenges. *Public Health Genomics* **18**, 242–247 (2015).
74. Helmy, M., Awad, M. & Mosa, K. A. Limited resources of genome sequencing in developing countries: challenges and solutions. *Appl. Transl Genom.* **9**, 15–19 (2016).
75. Satta, G., Atzeni, A. & McHugh, T. D. *Mycobacterium tuberculosis* and whole genome sequencing: a practical guide and online tools available for the clinical microbiologist. *Clin. Microbiol. Infect.* **23**, 69–72 (2017).
76. Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: scientific containers for mobility of compute. *PLOS ONE* **12**, e0177459 (2017).
77. Merkel, D. Docker: lightweight Linux containers for consistent development and deployment. *Linux J.* **2014**, 2 (2014).
78. Grüning, B. et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* **15**, 475–476 (2018).
79. Jackman, S., Birol, I., Jackman, S. & Birol, I. Linuxbrew and Homebrew for cross-platform package management. *F1000Res.* **5**, 1795 (2016).
80. Langille, M. G. I. & Eisen, J. A. BioTorrents: a file sharing service for scientific data. *PLOS ONE* **5**, e10071 (2010).
81. Karikari, T. K., Quansah, E. & Mohamed, W. M. Y. Widening participation would be key in enhancing bioinformatics and genomics research in Africa. *Appl. Transl Genom.* **6**, 35–41 (2015).
82. Bah, S. Y., Morang'a, C. M., Kengne-Ouafo, J. A., Amenga–Etego, L. & Awandare, G. A. Highlights on the application of genomics and bioinformatics in the fight against infectious diseases: challenges and opportunities in Africa. *Front. Genet.* **9**, 575 (2018).
83. Zignol, M. et al. Population-based resistance of *Mycobacterium tuberculosis* isolates to pyrazinamide and fluoroquinolones: results from a multicountry surveillance project. *Lancet Infect. Dis.* **16**, 1185–1192 (2016).
84. Kumwenda, S. et al. Challenges facing young African scientists in their research careers: a qualitative exploratory study. *Malawi Med. J.* **29**, 1–4 (2017).
85. Rabbani, F. et al. Schools of public health in low and middle-income countries: an imperative investment for improving the health of populations? *BMC Public Health* **16**, 941 (2016).
86. Helb, D. et al. Rapid detection of *Mycobacterium tuberculosis* and rifampin resistance by use of on-demand, near-patient technology. *J. Clin. Microbiol.* **48**, 229–237 (2010).
87. Wyllie, D. H. et al. Control of artifactual variation in reported intersample relatedness during clinical use of a *Mycobacterium tuberculosis* sequencing pipeline. *J. Clin. Microbiol.* **56**, e00104–18 (2018).
88. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
89. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).
90. Médigue, C., Cole, S. T., Camus, J.-C. & Pryor, M. J. Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology* **148**, 2967–2973 (2002).
91. Periwal, V. et al. Comparative whole-genome analysis of clinical isolates reveals characteristic architecture of *Mycobacterium tuberculosis* pangenome. *PLOS ONE* **10**, e0122979 (2015).
92. Gao, Q. et al. Gene expression diversity among *Mycobacterium tuberculosis* clinical isolates. *Microbiology* **151**, 5–14 (2005).
93. Kato-Maeda, M. et al. Comparing genomes within the species *Mycobacterium tuberculosis*. *Genome Res.* **11**, 547–554 (2001).
94. Alland, D. et al. Role of large sequence polymorphisms (LSPs) in generating genomic diversity among clinical isolates of *Mycobacterium tuberculosis* and the utility of LSPs in phylogenetic analysis. *J. Clin. Microbiol.* **45**, 39–46 (2007).
95. Ioerger, T. R. et al. Variation among genome sequences of H37Rv strains of *Mycobacterium tuberculosis* from multiple laboratories. *J. Bacteriol.* **192**, 3645–3653 (2010).
96. Lee, R. S. & Behr, M. A. Does choice matter? Reference-based alignment for molecular epidemiology of tuberculosis. *J. Clin. Microbiol.* **54**, 1891–1895 (2016).
97. Norman, A., Folkvardsen, D. B., Overballe-Petersen, S. & Lillebaek, T. Complete genome sequence of *Mycobacterium tuberculosis* DKC2, the predominant Danish outbreak strain. *Microbiol. Resour. Announc.* **8**, e01554–18 (2019).
98. Roetzer, A. et al. Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. *PLOS Med.* **10**, e1001387 (2013).
99. Bainomugisa, A. et al. A complete high-quality MinION nanopore assembly of an extensively drug-resistant *Mycobacterium tuberculosis* Beijing lineage strain identifies novel variation in repetitive PE/PPE gene regions. *Microb. Genom.* **4**, 256719 (2018).
100. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* **44**, 226–232 (2012).
101. Yadon, A. N. et al. A comprehensive characterization of *PncA* polymorphisms that confer resistance to pyrazinamide. *Nat. Commun.* **8**, 588 (2017).
102. Yang, Y. et al. Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. *Bioinformatics* **34**, 1666–1671 (2018).
103. Chen, M. L. et al. Deep learning predicts tuberculosis drug resistance status from genome sequencing data. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/275628v2 (2018).
104. Rajendran, V. & Sethumadhavan, R. Drug resistance mechanism of *PncA* in *Mycobacterium tuberculosis*. *J. Biomol. Struct. Dyn.* **32**, 209–221 (2013).
105. Kavvas, E. S. et al. Machine learning and structural analysis of *Mycobacterium tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance. *Nat. Commun.* **9**, 4306 (2018).
106. Duchêne, S. et al. Genome-scale rates of evolutionary change in bacteria. *Microb. Genom.* **2**, e000094 (2016).
107. Lee, R. S. et al. Reemergence and amplification of tuberculosis in the Canadian arctic. *J. Infect. Dis.* **211**, 1905–1914 (2015).
108. Clark, T. G. et al. Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing. *PLOS ONE* **8**, e83012 (2013).
109. Guthrie, J. L. et al. Genotyping and whole-genome sequencing to identify tuberculosis transmission to pediatric patients in British Columbia, Canada, 2005–2014. *J. Infect. Dis.* **218**, 1155–1163 (2018).
110. Bryant, J. M. et al. Whole-genome sequencing to establish relapse or re-infection with *Mycobacterium tuberculosis*: a retrospective observational study. *Lancet. Respir. Med.* **1**, 786–792 (2013).
111. Guerra-Assunção, J. A. et al. Recurrence due to relapse or reinfection with *Mycobacterium tuberculosis*: a whole-genome sequencing approach in a large, population-based cohort with a high HIV infection prevalence and active follow-up. *J. Infect. Dis.* **211**, 1154–1163 (2015).
112. Schürch, A. C. et al. The tempo and mode of molecular evolution of *Mycobacterium tuberculosis* at patient-to-patient scale. *Infect. Genet. Evol.* **10**, 108–114 (2010).
113. Lieberman, T. D. et al. Genomic diversity in autopsy samples reveals within-host dissemination of HIV-associated *Mycobacterium tuberculosis*. *Nat. Med.* **22**, 1470–1474 (2016).
114. Ford, C. B. et al. *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat. Genet.* **45**, 784–790 (2013).
115. Ford, C. B. et al. Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nat. Genet.* **43**, 482–486 (2011).
116. Hatherell, H.-A. et al. Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. *BMC Med.* **14**, 21 (2016). **This is a systematic review of the potential for WGS in determining transmission of MTBC strains**.
117. Verver, S. et al. Transmission of tuberculosis in a high incidence urban community in South Africa. *Int. J. Epidemiol.* **33**, 351–357 (2004).
118. Bjorn-Mortensen, K. et al. Tracing *Mycobacterium tuberculosis* transmission by whole genome sequencing in a high incidence setting: a retrospective population-based study in East Greenland. *Sci. Rep.* **6**, 33180 (2016).
119. Stimson, J. et al. Beyond the SNP threshold: identifying outbreak clusters using inferred transmissions. *Mol. Biol. Evol.* **36**, 587–603 (2019).
120. Biek, R., Pybus, O. G., Lloyd-Smith, J. O. & Didelot, X. Measurably evolving pathogens in the genomic era. *Trends Ecol. Evol.* **30**, 306–313 (2015).
121. Campbell, F. et al. outbreaker2: a modular platform for outbreak reconstruction. *BMC Bioinformatics* **19**, 363 (2018).
122. Didelot, X., Gardy, J. & Colijn, C. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol. Biol. Evol.* **31**, 1869–1879 (2014).
123. Didelot, X., Fraser, C., Gardy, J. & Colijn, C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Mol. Biol. Evol.* **34**, 997–1007 (2017).
124. De Maio, N., Worby, C. J., Wilson, D. J. & Stoesser, N. Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLOS Comput. Biol.* **14**, e1006117 (2018).
125. Klinkenberg, D., Backer, J. A., Didelot, X., Colijn, C. & Wallinga, J. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLOS Comput. Biol.* **13**, e1005495 (2017).
126. Kühnert, D. et al. Tuberculosis outbreak investigation using phylodynamic analysis. *Epidemics* **25**, 47–53 (2018).
127. Eldholm, V. et al. Armed conflict and population displacement as drivers of the evolution and dispersal of *Mycobacterium tuberculosis*. *Proc. Natl Acad. Sci. USA* **113**, 13881–13886 (2016).
128. Streicher, E. M. et al. *Mycobacterium tuberculosis* population structure determines the outcome of genetics-based second-line drug resistance testing. *Antimicrob. Agents Chemother.* **56**, 2420–2427 (2012).
129. Folkvardsen, D. B. et al. Rifampin heteroresistance in *Mycobacterium tuberculosis* cultures as detected by phenotypic and genotypic drug susceptibility test methods. *J. Clin. Microbiol.* **51**, 4220–4222 (2013).
130. Shamputa, I. C. et al. Mixed infection and clonal representativeness of a single sputum sample in tuberculosis patients from a penitentiary hospital in Georgia. *Respir. Res.* **7**, 99 (2006).
131. Sobkowiak, B. et al. Identifying mixed *Mycobacterium tuberculosis* infections from whole genome sequence data. *BMC Genomics* **19**, 613 (2018).
132. Gan, M., Liu, Q., Yang, C., Gao, Q. & Luo, T. Deep whole-genome sequencing to detect mixed infection of *Mycobacterium tuberculosis*. *PLOS ONE* **11**, e0159029 (2016).
133. Votintseva, A. A. et al. Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples. *J. Clin. Microbiol.* **55**, 1285–1298 (2017).

134. Doyle, R. M. et al. Direct whole-genome sequencing of sputum accurately identifies drug-resistant *Mycobacterium tuberculosis* faster than MGIT culture sequencing. *J. Clin. Microbiol.* **56**, e00666–18 (2018).
135. Doughty, E. L., Sergeant, M. J., Adetifa, I., Antonio, M. & Pallen, M. J. Culture-independent detection and characterisation of *Mycobacterium tuberculosis* and *M. africanum* in sputum samples using shotgun metagenomics on a benchtop sequencer. *PeerJ* **2**, e585 (2014).
136. Phelan, J. E. et al. Recombination in pe/ppe genes contributes to genetic variation in *Mycobacterium tuberculosis* lineages. *BMC Genomics* **17**, 151 (2016).
137. Reisner, B. S., Gatson, A. M. & Woods, G. L. Evaluation of mycobacteria growth indicator tubes for susceptibility testing of *Mycobacterium tuberculosis* to isoniazid and rifampin. *Diagn. Microbiol. Infect. Dis.* **22**, 325–329 (1995).
138. Strydom, K. et al. Comparison of three commercial molecular assays for detection of rifampin and isoniazid resistance among *Mycobacterium tuberculosis* isolates in a high-HIV-prevalence setting. *J. Clin. Microbiol.* **53**, 3032–3034 (2015).
139. Nathavitharana, R. R. et al. Multicenter noninferiority evaluation of Hain GenoType MTBDRplus version 2 and Nipro NTM+MDRTB line probe assays for detection of rifampin and isoniazid resistance. *J. Clin. Microbiol.* **54**, 1624–1630 (2016).
140. Mitarai, S. et al. Comprehensive multicenter evaluation of a new line probe assay kit for identification of Mycobacterium species and detection of drug-resistant *Mycobacterium tuberculosis*. *J. Clin. Microbiol.* **50**, 884–890 (2012).
141. Hillemann, D., Rüsch-Gerdes, S. & Richter, E. Feasibility of the GenoType MTBDRsl assay for fluoroquinolone, amikacin-capreomycin, and ethambutol resistance testing of *Mycobacterium tuberculosis* strains and clinical specimens. *J. Clin. Microbiol.* **47**, 1767–1772 (2009).
142. Tagliani, E. et al. Diagnostic performance of the new version (v2.0) of GenoType MTBDR *sl* assay for detection of resistance to fluoroquinolones and second-line injectable drugs: a multicenter study. *J. Clin. Microbiol.* **53**, 2961–2969 (2015).
143. Ng, K. C. et al. Potential application of digitally linked tuberculosis diagnostics for real-time surveillance of drug-resistant tuberculosis transmission: validation and analysis of test results. *JMIR Med. Inform.* **6**, e12 (2018).
144. Chakravorty, S. et al. The new Xpert MTB/RIF Ultra: improving detection of *Mycobacterium tuberculosis* and resistance to rifampin in an assay suitable for point-of-care testing. *mBio* **8**, e00812–17 (2017).
145. Ng, K. C. S. et al. Xpert Ultra can unambiguously identify specific rifampin resistance-conferring mutations. *J. Clin. Microbiol.* **56**, e00686–18 (2018).
146. Molina-Moya, B. et al. Diagnostic accuracy study of multiplex PCR for detecting tuberculosis drug resistance. *J. Infect.* **71**, 220–230 (2015).
147. Hillemann, D., Haasis, C., Andres, S., Behn, T. & Kranzer, K. Validation of the FluoroType MTBDR assay for detection of rifampin and isoniazid resistance in *Mycobacterium tuberculosis* complex isolates. *J. Clin. Microbiol.* **56**, e00072–18 (2018).
148. Pang, Y. et al. Rapid diagnosis of MDR and XDR tuberculosis with the MeltPro TB assay in China. *Sci. Rep.* **6**, 25330 (2016).
149. Kaswa, M. K. et al. Pseudo-outbreak of pre-extensively drug-resistant (Pre-XDR) tuberculosis in Kinshasa: collateral damage caused by false detection of fluoroquinolone resistance by GenoType MTBDRsl. *J. Clin. Microbiol.* **52**, 2876–2880 (2014).
150. Ajileye, A. et al. Some synonymous and nonsynonymous *gyrA* mutations in *Mycobacterium tuberculosis* lead to systematic false-positive fluoroquinolone resistance results with the Hain GenoType MTBDRsl assays. *Antimicrob. Agents Chemother.* **61**, e02169–16 (2017).
151. Colman, R. E. et al. Detection of low-level mixed-population drug resistance in *Mycobacterium tuberculosis* using high fidelity amplicon sequencing. *PLOS ONE* **10**, e0126626 (2015).
152. Colman, R. E. et al. Rapid drug susceptibility testing of drug-resistant *Mycobacterium tuberculosis* isolates directly from clinical samples by use of amplicon sequencing: a proof-of-concept study. *J. Clin. Microbiol.* **54**, 2058–2067 (2016).
153. Makhado, N. A. et al. Outbreak of multidrug-resistant tuberculosis in South Africa undetected by WHO-endorsed commercial tests: an observational study. *Lancet Infect. Dis.* **18**, 1350–1359 (2018).
154. Tagliani, E. et al. Culture and next-generation sequencing-based drug susceptibility testing unveil high levels of drug-resistant-TB in Djibouti: results from the first national survey. *Sci. Rep.* **7**, 17672 (2017).
155. Gagneux, S. Ecology and evolution of *Mycobacterium tuberculosis*. *Nat. Rev. Microbiol.* **16**, 202–213 (2018).

Acknowledgements

C.J.M. and L.R. are also affiliated with BCCM/ITM Mycobacterial Culture Collection, Institute of Tropical Medicine, Antwerp, Belgium. J.G. is also affiliated with the BC Centre for Disease Control, Vancouver, Canada. B.O.-A. is also affiliated with the Center for Global Health Security and Diplomacy, Ottawa, Canada. M.S. is also affiliated with the University of Arizona, Tucson, AZ, USA. I.C. is also affiliated with the CIBER in Epidemiology and Public Health, Spain. C.J.M., B.O.-A., L.R. and B.C.d.J. are supported by a European Research Council grant (INTERRUPTB; no. 311725). I.C. and G.A.G. are supported by a European Research Council grant (TB-ACCELERATE; no. 638553). T.C.R. receives salary support from the not-for-profit organization Foundation for Innovative New Diagnostics (the terms of this arrangement have been reviewed and approved by the University of California, San Diego). T.M.W. is an NIHR Academic Clinical Lecturer. J.L.G. and J.G. receive funding from the University of British Columbia, Vancouver, Canada. T.A.K., C.U., V.D. and S.N. receive funding from the German Center for Infection Research (DZIF) and are funded by the Deutsche Forschungsgemeinschaft (German Research Foundation) under Germany's Excellence Strategy (EXC 22167–390884018). L.V., T.H.H. and A.V.R. are funded by FWO Odysseus G0F8316N. M.R.F. is supported by the US National Institutes of Health BD2K K01 (MRF ES026835). P.S. is supported by the Agence Nationale de la Recherche (ANR-16-CE35-0009).

Author contributions

C.J.M., G.A.G., T.A.K., L.V., A.D., M.E., M.R.F., J.L.G., K.L., P.M., B.O.-A., V.D., P.S., A.S., C.U., D.v.S., Y.Z., M.S., J.G., D.M.C., S.N., I.C. and A.V.R. researched the data for the article. C.J.M., G.A.G., T.A.K., L.V., A.D., M.E., M.R.F., J.L.G., K.L., P.M., B.O.-A., V.D., P.S., A.S., C.U., D.v.S. Y.Z., M.d.V., S.G., T.H.H., L.R., E.T., T.M.W., R.M.W., M.S., J.G., D.M.C., S.N., I.C. and A.V.R. substantially contributed to the discussion of the content. C.J.M., G.A.G., T.A.K., L.V., A.D., M.E., M.R.F., J.L.G., K.L., P.M., B.O.-A., V.D., P.S., A.S., C.U., D.v.S., Y.Z., M.S., J.G., D.M.C., S.N., I.C. and A.V.R. wrote the article. All authors reviewed and edited the manuscript before submission.

Competing interests

P.S. was a consultant for Genoscreen. All other authors declare no competing interests.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Reviewer information

Nature Reviews Microbiology thanks T. McHugh, V. Sintchenko, and other anonymous reviewer(s), for their contribution to the peer review of this work.

Supplementary information

Supplementary information is available for this paper at https://doi.org/10.1038/s41579-019-0214-5.

RELATED LINKS

Human, Heredity and Health in Africa Consortium: https://h3abionet.org
ReSeqTB: http://www.reseqtb.org
TORCH consortium: https://torch-consortium.com/vliruos
ERLTB-Net: https://ecdc.europa.eu/en/about-us/partnerships-and-networks/disease-and-laboratory-networks/erltb-net