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

Conor J Meehan, Galo A. Goig, Thomas Andreas Kohl, Lennert Verboven, Anzaan Dippenaar, Matthew Ezewudo, Maha Farhat, Jennifer L. Guthrie, Kris Laukens, Paolo Miotto, Boatema Ofori-Anyinam, Viola Dreyer, Philip Supply, Anita Suresh, Christian Utpatel, Dick van Soolingen, Yang Zhou, Philip Ashton, Daniela Brites, Andrea M. Cabibbe, Bouke C. de Jong, Margaretha de Vos, Fabrizio Menardo, Sebastien Gagneux, Qian Gao, Tim H Heupink, Qingyun Liu, Chloé Loiseau, Leen Rigouts, Timothy C Rodwell, Elisa Tagliani, Timothy M. Walker, Robin Mark Warren, Yanlin Zhao, Matteo Zignol, Marco Schito, Jennifer Gardy, Daniela Maria Cirillo, Stefan Niemann, Inaki Comas* and Annelies Van Rie*

Affiliations:

CJM Unit of Mycobacteriology, Department of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium

GAG Institute of Biomedicine of Valencia, CSIC and CIBER in Epidemiology and Public Health, Valencia, Spain

TAK Molecular and Experimental Mycobacteriology, Priority Area Infections, Research Center Borstel, D-23845 Borstel, Germany

LV Tuberculosis Omics Research Consortium, Department of Epidemiology and Social Medicine, Institute of Global Health, Faculty of Medicine and Health Sciences, University of Antwerp

AD 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

ME Critical Path Institute, Arizona, USA

MF Harvard Medical School, and Massachusetts General Hospital, Boston, MA, USA

JLG University of British Columbia, Vancouver, Canada

KL Adrem Data Lab, Department of Mathematics & Computer Science, University of Antwerp

PM Emerging Bacterial Pathogens Unit, Division of Immunology, Transplantation and Infectious Diseases, IRCCS San Raffaele Scientific Institute, Milano, Italy

BO Center for Global Health Security and Diplomacy, Ottawa, Canada, Food and Drugs Authority, Accra, Ghana

VD Molecular and Experimental Mycobacteriology, Priority Area Infections, Research Center Borstel, D-23845 Borstel, Germany

PS University Lille, CNRS, INSERM, CHU Lille, Institut Pasteur de Lille, U1019, UMR 8204, CIIL, Centre d'Infection et d'Immunité de Lille, Lille, France

AS Foundation for Innovative New Diagnostics (FIND), Geneva, Switzerland

CU Molecular and Experimental Mycobacteriology, Priority Area Infections, Research Center Borstel, D-23845 Borstel, Germany

DvS National Tuberculosis Reference Laboratory, Centre for Infectious Disease Control, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands

ZY National Center for Tuberculosis Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China

PA Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, UK

DB Swiss Tropical and Public Health Institute, Basel, Switzerland

AMC Emerging Bacterial Pathogens Unit, Division of Immunology, Transplantation and Infectious Diseases, IRCCS San Raffaele Scientific Institute, Milano, Italy

BCdJ Unit of Mycobacteriology, Department of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium

MDV 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

FM Swiss Tropical and Public Health Institute, Basel, Switzerland
SG Swiss Tropical and Public Health Institute, Basel, Switzerland
QG Key Laboratory of Medical Molecular Virology, Ministry of Education and Health, School of Basic Medical Sciences, Fudan University, Shanghai, China.
THH Tuberculosis Omics Research Consortium, Department of Epidemiology and Social Medicine, Institute of Global Health, Faculty of Medicine and Health Sciences, University of Antwerp
CL Swiss Tropical and Public Health Institute, Basel, Switzerland
QL Key Laboratory of Medical Molecular Virology, Ministry of Education and Health, School of Basic Medical Sciences, Fudan University, Shanghai, China.
LR Unit of Mycobacteriology, Department of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium
TCR Foundation for Innovative New Diagnostics (FIND), Geneva, Switzerland
ET Emerging Bacterial Pathogens Unit, Division of Immunology, Transplantation and Infectious Diseases, IRCCS San Raffaele Scientific Institute, Milano, Italy
TMW Nuffield Department of Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK
RMW 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
YZ National Center for Tuberculosis Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China
MZ Global Tuberculosis Programme, World Health Organization, Geneva, Switzerland
MS Critical Path Institute, Arizona, USA
JG University of British Columbia, Vancouver, Canada
DMC Emerging Bacterial Pathogens Unit, Division of Immunology, Transplantation and Infectious Diseases, IRCCS San Raffaele Scientific Institute, Milano, Italy
SN Molecular and Experimental Mycobacteriology, Priority Area Infections, Research Center Borstel, D-23845 Borstel, Germany
IC Institute of Biomedicine of Valencia, CSIC and CIBER in Epidemiology and Public Health, Valencia, Spain
AVR Tuberculosis Omics Research Consortium, Department of Epidemiology and Social Medicine, Institute of Global Health, Faculty of Medicine and Health Sciences, University of Antwerp

Abstract

Whole genome sequencing (WGS) of *Mycobacterium tuberculosis* has rapidly evolved from a research tool to a clinical application for the diagnosis and management of tuberculosis and in public health surveillance. This evolution has been facilitated by the dramatic drop in costs, 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 harmonisation, 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 repository of resistance-causing variants, phylogenetic analyses, quality control processes, and standardised reporting.

1. **Introduction**

Mycobacterium tuberculosis complex (Mtbc) pathogens are collectively the top infectious disease killer globally, causing 10 million new tuberculosis (TB) cases annually[1]. Increasingly, new TB cases are already resistant to rifampicin and isoniazid (termed multidrug resistance; MDR-TB), the key first line drugs[1]. Tackling the spread and drug resistance burden of this pathogen 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 *Mtb* transmission dynamics[2,3].

Whole genome sequencing (WGS) approaches use DNA sequencing platforms to reconstruct the complement of DNA found inside a cell. The small (~4.4Mb), single chromosome genome of Mtbc strains[4] lends itself well to WGS approaches. . Rapid, reliable, and increasingly affordable WGS technologies, can now guide all components of TB control: diagnosis, treatment, surveillance and contact tracing[5,6] (Fig. 1). Individual (sub)species of human and animal Mtbc lineages can be identified,[7–9] and drug resistance profiles can be predicted, especially well for 1st line drugs[2], enabling prompt, appropriate initiation of treatment and monitoring the acquisition of drug resistance[10]. TB outbreaks can be identified with high resolution[11–13], including across borders,[14,15] and diseases control measures implemented. The analysis of the emergence, spread, genetic makeup, and evolution of particular outbreak strains, e.g. highly resistant or highly virulent clones, can enable the development of targeted measures[16–18].

WGS-based approaches are quickly moving from research-only to clinical care and public health applications. The World Health Organization (WHO) is already using WGS for drug resistance surveillance[19] and is scheduled to evaluate sequencing technologies for routine genotypic drug susceptibility testing in 2019[1]. As WGS-guided individualized treatment[20] and WGS-based surveillance systems[15] are being implemented in several countries (e.g. the UK and the Netherlands) with more to come, accurate methods and standardized reporting are vital. At present, multiple WGS data analysis solutions exist that vary widely in scope, pipelines, and output formats, with little standardisation amongst them[21], making cross-comparisons and rigorous validation of these pipelines difficult. Because clinical decisions such as the effective drugs that can be included in a patients' regimen may be influenced by differences in the bioinformatic analysis, robustness of the pipeline used in clinically-relevant predictions tools is critical.

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/lineage identification (referred to as strain typing). We highlight those places where a general agreement in the analysis parameters or interpretation of the results has been already reached by the community. Alternatively, we discuss those items where there is still open discussion about the best practices and will require more effort to reach a consensus in the future.

2. State of the art

The standard workflow for WGS analysis of Mtbc strains is outlined in Figure 2. It involves culturing sputum specimens on solid (Löwenstein–Jensen) or liquid (Mycobacteria Growth Indicator Tube) media, extracting DNA from Mtbc strains, library preparation, and sequencing by short read technologies (e.g. Illumina platforms)[22]. 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 H37Rv) and detection of genomic variants such as single nucleotide polymorphisms (SNPs) and insertion/deletions (indels). Numerous resequencing pipelines for the Mtbc currently exist with currently no single ‘gold standard’. These pipelines typically exclude about ~10% of the genome because erroneous mapping in certain regions result in false variant calls (PE/PPE gene families, other repetitive genes, mobile elements[4]) and apply various criteria, such as read depth, base quality, and strand bias to filter out false positive variants. Finally, based on the detected variants, several tasks can be performed including (but not limited to) prediction of drug resistance and susceptibility profiles, strain typing, and identification of transmission clusters.

Due 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[23,24]. Drug resistance and susceptibility profiles can be determined with high accuracy for many drugs used for the treatment of TB by comparing variant calls to 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[25,26] (Fig. 3). A prime effort in the construction of these lists is the Relational Sequencing Tuberculosis Data Platform (ReSeqTB, http://www.reseqtb.org), where researchers from around the world can contribute data[27]. 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 (MIC) for most drugs used for TB treatment[2]. By comparing the SNPs present in a sequenced isolate to these lists, WGS can not only predict resistance but also 1st line pan-susceptibility 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 sub-lineages, 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[28] and comparing isolates on the global level[18,29].

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 although genome-based multi locus sequence typing (MLST) has shown comparable results[30,31]. 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[32]. It is not yet clear if a single threshold can be employed to detect epidemiologically linked cases in all timeframes and contexts. The MLST approach employs 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 schema exist for this approach: the core genome (termed cgMLST; 2891 genes covering 2.86 million bases[31]) and an extended pan-genome including 1141 accessory loci[11] (termed wgMLST). These WGS-based approaches have been shown to perform better than contact tracing and with higher resolution than classical approaches such as MIRU-VNTR[12,13,30,31,33].

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 standardisation and validation of this workflow are required to ease integration into current clinical and public health initiatives.

Currently, two Mtbc-specific pipelines are available, which perform multiple core tasks in single install set-up to produce genetic variant calls from raw Illumina sequence data (MTBseq[34] and UVP-ReSeqTB[35]). Other pathogen-agnostic pipelines can be used with an Mtbc-specific reference genome and drug resistance database to achieve similar results[33,36–38]. Numerous custom-built pipelines also exist[8,39–46], 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 routinely implemented. Lastly, pipelines specific for a single task such as drug resistance prediction[25,47–51] or strain typing[7,50] are available and have been comprehensively compared elsewhere[52–55].

3. Mtbc WGS validation and standardisation

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 (standardisation). Ideally, all steps of the workflow, from DNA extraction to sequencing, data analysis and reporting, should be standardised (or at least comparable) and well documented, and an external quality assessment (EQA) program should be in place. Efforts to standardise and validate the upstream (pre-bioinformatics pipeline) steps have been undertaken to great effect[22,54]. Pipeline standardisation could be achieved through the use of a single pipeline in all settings or through validation with rigorous testing and convergence on a defined outcome for all pipelines developed. Since multiple pipelines have already been implemented (e.g. MTBseq[34] for the EUSeqMyTB consortium and the Unified Variant Pipeline[35] 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 standardise these multiple pipelines for clinical use[56]. In 2016, the US Food and Drug Administration (FDA) released draft guidelines on sequencing-based infectious disease diagnostics and bodies such as the WHO and ECDC are taking steps towards international standardisations of Mtbc WGS[15,22,57].

a. Technical validation and external quality control of Mtbc WGS

First, the extraction of DNA needs to meet minimal standards as defined for a given WGS instrument[22]. Next, the pipeline to convert the raw sequencing reads into accurate variant calls should be technically valid, i.e. call the correct variantss. While there is much debate about the

reference standard to be used for technical validation of WGS pipelines, currently this is best undertaken by using short read datasets derived from isolates with known complete genomes (e.g. from long read sequencing)[58]. Mapping these read sets to their respective assembled genomes allows to calculate the rate of false positive and negative SNPs called by the pipeline under consideration. Ideally, to promote interoperability and ease the verification of bioinformatics protocols, a standard reporting format such as a BioCompute Object (BCO) to record all thresholds, steps and implementation arguments for a given pipeline is utilised[59]. 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[60].
A prime example of external quality control of bioinformatics pipelines is the efforts by the 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)[21]. Panels of DNA extracted from selected Mtbc isolates are sent annually by RIVM to reference laboratories to assess intra- and inter laboratory reproducibility of WGS. Similar efforts in high burden settings are needed to monitor the reliability of Mtbc WGS outputs when used in these settings.

b. 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, e.g. drug resistance profiling. For task validation, Mtbc bioinformatics pipelines should use defined validation datasets, ideally with hundreds or thousands of well characterized clinical Mtbc strains representing the diversity of a specific core task (e.g. different drug susceptibility profiles for resistance detection, representatives of all Mtbc lineages and (sub)-species for typing, or varying degrees of clustering for transmission analyses). The number of readily available, well-curated validation datasets is currently limited.

Validation of transmission clustering. The national public health institute of the Netherlands (RIVM) has provided laboratories with sequenced reads from 535 Mtbc isolates for which epidemiological links were known. Using this dataset, 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 standardise WGS for monitoring MDR-TB cross border transmission in Europe[15].

Validation of classification systems. 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 (sub)lineage[8,9,61]. However, sub-lineage classifications are often less resolved, and parallel nomenclatures for lineage 2 are being used[18,62,63]. As the diversity of the Mtbc is further explored, especially for animal-associated and zoonotic TB, these under-described lineages can also easily be typed using the same SNP-based approach[7].

Validation of drug resistance profiling**.** Validation of WGS for TB resistance is the most advanced of all the core tasks. Phelan *et al* showed high concordance between phenotypic and genotypic predictions, no matter the sequencing platform used[19,54]. In the past two years, major progress has been made in the linkage between genotype and resistance phenotype by employing a standardized statistical approach[25,26]. The task of incrementally improving our knowledge base on genetic resistance profiling is primarily being addressed by the two global consortia outlined above: ReSeqTB's single platform for genotype-phenotype investigation of drug resistance[27,35] and CRyPTIC's genotypic-phenotypic linking of over 10,000 isolates demonstrating susceptibility prediction for rifampicin and isoniazid with 99% sensitivity and 93-96% for ethambutol and pyrazinamide[2]. These results have led to some low burden countries (Netherlands, UK) replacing phenotypic DST with WGS-based DST for first line drugs. Resistance predictions for 2nd line drugs can also be undertaken with sensitivity often around 90%[25]. Large comparative studies using phenotype-genotype associations are expanding the catalogues[64,65] and will help to increase the sensitivity for drugs used to treat MDR-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.

a. Standardization of communication of Mtbc WGS results and data sharing

Communication to end users: Effective communication of WGS-based results to a diverse audience of end-users is key to positively impacting patient care and TB control programs. While the need for plain language reporting of genomic results has been recognized[52,66], there are no international standards yet. Reporting standards should be flexible enough to address the varying levels of familiarity of end-users with genomic data interpretation and allow customization to region-specific treatment guidelines and formatting requirements. For example, the ISO15189:2012 standard mandates information such as patient identifiers, assay details, and the testing laboratory be reported. Recommendations from Mtbc WGS report design validation studies included the use of complete terms instead of abbreviations, drawing attention to important elements with shading, bolding, and other types of emphasis, and incorporating summary statements to rapidly communicate key results[67,68].

Communication to the research community. In peer-reviewed publications, the parameters used at each step of a bioinformatics pipeline must be stated in a way that makes it reproducible and understandable to non-bioinformaticians (e.g. using a BCO as outlined above). Custom code used in the analysis should be made available through a public repository (e.g.GitHub), ensuring ease of installation elsewhere. Pipelines should report the outcome of technical validations, at least for the core tasks they aim to address (e.g. lineage-defining SNPs for a typing pipeline). Examples of standard reporting include the MIABi (Minimum Information About a Bioinformatics investigation)[69] and the STROME-ID (Strengthening the Reporting of Molecular Epidemiology for Infectious Diseases)[70] guidelines. In supplementary table 2, 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 will be crucial as incremental knowledge improves drug resistance predictions and strain tracking relies on the number and diversity of strain genome data available. This can come in the form of sharing coded strain identifiers such as MLST patterns or raw sequence data not yet processed by a pipeline. Indeed data sharing has been shown already 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 surveillance efforts[71], similar to the GenomeTrakr network for foodborne pathogens[72], while still safeguarding patient data and appropriately acknowledging contributions. This setup would be of great value for moving the field of Mtbc WGS forward.

The crucial next step for fully utilising Mtbc WGS data is implementation of validations, both technical and task oriented, for all pipelines. Once undertaken, the agreed upon pipeline(s) can then be widely implemented, once infrastructure and usability is accounted for.

2. **Implementation of WGS in routine clinical practice**

While the use of WGS is rapidly expanding in research, minimal progress has been made in programmatic use of WGS. Some reasons include the lack of standardised end-to-end solutions, the required wet-lab and computing infrastructure, need for sufficient internet connectivity and bandwidth, and training deficits in genomics and bioinformatics[73–75]. Efforts are thus needed to expand accessibility to perform analysis by non-experts. How these factors are addressed will depend 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 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 cloud/web-based approach with easy, affordable access[76]. Many large healthcare facilities such as referral hospitals are already incorporating bioinformatics units into their support services as part of the push towards personalized medicine, something 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[34,77].

Giving the heterogeneity of pipelines already in place (e.g. supplementary table 1) it is conceivable that something similar will happen when implementation is done in hundreds of care services. Some will opt-in for end-to-end solutions, perhaps integrated with the sequencing platform, or others for task-specific, 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[6,76]. In order to evaluate new pipelines it is preferable to develop inside 'containers', such as Docker or Singularity[78,79], or one-command installation wrappers like Bioconda or Homebrew[80,81]. Creating a container for each step (Figure 2) also allows for easy updating of a specific step without the need to install a whole new pipeline and allows for tasks (e.g. resistance profiling) to be added to the pipeline as needed. To allow usability by a range of end-users, fine-grained access to the individual steps should be available for advanced users with functionality layers abstracted away for users with limited bioinformatics expertise. The pipelines should be open source and user-friendly, by employing intuitive and well-documented command line and graphical user interfaces with relevant and validated default parameters.

The situation in 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 qPCR systems (Box 1). Centralized web-based analysis platforms have recently emerged and promise to aid in computational efficiency, access and usability[47,51]. Roll-out of such initiatives to more countries would greatly improve the potential for large-scale WGS implementation. The primary barrier to this is usually unstable internet connectivity with limited bandwidth, although using methods that can effectively handle connection interruptions, such as BioTorrents[82], 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 play an important role in LMICs. It is, however, advisable to build in those countries human capacity for WGS of Mtbc strains[83,84]. While standardised, immutable 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 customised, yet reproducible solutions are being supported by capacity building initiatives (e.g. the Human, Heredity and Health in Africa Consortium (https://h3abionet.org) and the TORCH consortium (https://torch-consortium.com/vliruos)). TB supranational reference laboratories should also play an important coordinating role, as is currently done for phenotypic workflows[19,85]. Ultimately, expanding education curricula to include bioinformatics are needed to generate sufficient capacity[86].

Finally, supportive policy and political commitment will be essential for sustainable implementation of WGS, especially in TB endemic LMICs[74,83,87]. This implementation will benefit from the lessons learned during the step-wise approach used to roll-out line probe assays and GeneXpert (Box 1)[88].

3. Extensions of the current standard

While 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.

a. Input data validation and quality control

Most current pipelines do not routinely filter out reads that do not come from the sequenced 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[89]. Additionally, any host DNA sequencing reads should be removed especially if the data is shared online for legal/ethical reasons. Computationally removing non-Mtbc strain reads prior to mapping is an efficient strategy to implement contamination-proof analysis pipelines[40], but requires a taxonomic classification of individual reads. Using taxonomic classification methods, where reads are assigned to the closest matched species, allows for quick and efficient removal of contaminating reads but requires comprehensive genome databases, often making their implementation extremely memory consuming[90,91]. Additionally, elimination of reads from highly conserved core bacterial genes of heterologous sources still remains a problem. Proposed alternatives include masking genomic regions known to accumulate artefactual polymorphisms[89], 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 methodology will require thorough technical validation to ensure that contaminant reads are removed without eliminating true Mtbc sequences, e.g. through *in silico* generation of datasets with varying levels of reads from other organisms.

b. 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. While most pipelines use the H37Rv genome[4,92] as the reference genome, several alternative approaches should be explored. Since H37Rv is a lineage 4 strain, its use as a reference for other lineages may be insufficient due to gene content differences between lineages[93–96]. Additionally, H37Rv contains many variants not found in any other strain[97], including in genes related to drug resistance (e.g. *gyrA*S95T), creating confusion in SNP interpretations. Any replacement of H37Rv as the reference genome should be assessed by *in-silico* studies across datasets and clinical settings. An example of such a study tested seven different references against sequence reads from lineage 4 isolates showing that very limited variation occurred, and that reference choice should be based on criteria other than matching lineage[98].

One alternative to the H37Rv genome is a pan-genome which incorporates the entire gene pool of *Mtbc* lineages. Previous studies have found small but notable differences in gene content between lineages, often affecting genes involved in pathogenesis[93–96]. While these differences are unlikely to affect drug resistance profiling (since 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 a *Mtbc* pan-genome should be straightforward due to the close 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[29,40]. From an evolutionary perspective, this approach addresses the H37Rv-specific variants outlined above. In addition, because all extant strains are equidistant to a common ancestor, the number 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, co/super-infections and contaminations[40].

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 genome[99–101] could be used as reference to reduce mapping errors[10]. A disadvantage of this approach is that it hampers comparison of results between pipelines and 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[102].

Independent of the selection of the reference genome, other steps such as mapping and filtering are not consistent between different pipelines, yet might greatly affect the analysis outcome. For instance, removal of duplicates, both PCR and optical, may have a large impact in the variants identified and the allele frequencies. Similarly, local assembly/realignment around indels, reducing false positive SNPs derived from mapping artefacts, is rarely used in Mtbc WGS pipelines[58] but is known to affect variant calling[47]. The question of whether these steps have a relevant effect on the final outcome should be incorporated into future technical validations.

c. Interpretation of drug resistance results and predictions

Currently, the bulk of routine drug resistance testing is undertaken using pDST. While this approach will still be required for a subset of difficult to interpret drug resistance patterns, the overarching goal is to detect all variants associated with resistance for comprehensive genome-based resistance profiling. While the current statistical approach to calling 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 new and repurposed drugs, or drugs such as pyrazinamide and ethionamide for which mutations are not limited to hotspots but appear across genes (*pncA* and *ethA)* and in promoter regions. For these drugs, the standard statistical approach could be complemented by experimental data, comprehensive single nucleotide mutagenesis[103] followed by systematic phenotypic screening, multi-omics studies, and machine learning approaches to predict the resistance phenotype of uncommon or novel genomic variants[104,105]. With the final aim of replacing the majority of phenotypic DST by sequence-based testing, it will also be essential to catalogue “benign” variants that are not associated to resistance, i.e. phylogenetic markers or other neutral variants[2]. New statistical approaches like large-scale GWAS[64,65], protein structure modelling[44,106] and machine learning[104,105,107] will likely play a key role identifying causative versus benign variants. Comprehensive databases of WGS data linked with phenotypic and clinical outcome data (e.g. CRyPTIC or ReSeqTB) are key to moving towards this goal.

Once established, endorsement of a single standardised variant list by the WHO or other regulating bodies, with regular updating should be favoured.

d. Variant calling for other purposes

Accurate variant calling has major implications on downstream interpretation of the results for evolutionary, epidemiological and clinical applications. Because of the low level diversity and the slow substitution rate of Mtbc genomes[32,42,100,108], a few falsely called SNPs can affect the interpretation of transmission events, impact the classification of a second episode of TB as relapse versus re-infection, or influence the interpretation of sub-populations within a patient (Fig. 4).

A primary use of Mtbc WGS is the identification of recent transmission chains and its direction at high resolution. While some studies have used thresholds from 0 to <50 SNPs[109–111], a threshold of 5- or 12-SNP genetic distances is most frequently used to identify possible epidemiological links and recent transmission[30,32]. For WGS-based distinction of relapse versus reinfection, studies have used often arbitrary thresholds of < 6 or <10 SNPs to define reactivation, and >100 to >1306 to define re-infection[46,112,113]. 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[42,114], clonal variants in different lesions[10,115], the impact of strain type (lineage or subspecies) or drug resistance on substitution rates[108,116], and genome stability/instability during latency[116,117]. 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[118–120].

Other approaches have meanwhile been developed to improve the identification of epidemiological links and outbreak reconstruction beyond SNP-based clustering. These either use transmission event thresholds[121] and/or often combine genomic and epidemiological data to identify the most probable transmission trees for infectious diseases[122,123]. Of particular importance when reconstructing Mtbc outbreaks is that phylogeny and transmission events do not necessarily coincide as a results of genetic diversification during latency and long generation times[124]; it is thus necessary to model the within-host genetic dynamics[125–127]. Besides transmission reconstruction, phylodynamic approaches also allow for the inference of epidemiological relevant parameters such as the effective reproduction number as well as the timing and geographic origin of an outbreak[128,129].

Unravelling within-host dynamics in terms of subpopulation detection remains even more challenging. Low frequency variants that are not due to technical artefacts can indicate the presence of mixed infections (two distinct Mtbc strains co-circulating in a host), or microevolution leading to closely related subpopulations or heteroresistance (subpopulations that differ in drug resistance-related variants)[10,115,130]. Proposed sub-population detection limits in different pipelines vary considerably from 10% to <75% (supplementary table 1) and are strongly influenced by factors such as read depth. While the presence of a subpopulation of at least 1% resistant bacilli is considered clinically relevant[131] , the chain reaction of 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[115,132]. Mathematical modelling approaches have been developed to identify mixed infections[133,134]. However with the current approaches the detection of mixed infections is limited by the relative ratio of the two strains and the number of differing SNPs between both. Future research and methodological improvements are needed to better understand and interpret this within-host diversity.

4. Beyond the current standards

As current culture-based approaches require time for Mtbc strain growth, culture-free WGS, directly from clinical samples (e.g.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. While 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[135–137]. 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 (e.g. the MinION). Such technology offers the capacity to detect variants in real-time during sample acquisition, potentially giving 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 LMICs. Moreover, although progress has been made in analysis of variants in repeat-rich genome regions (e.g. PE/PPE family genes) or structural changes (duplications, large indels, etc.) by short read mapping[112,138], long read sequencing will make this more robust[101,135]. 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 over-shearing DNA.

5. **Conclusion**

A decade after first proof-of-principle studies, the community consensus is that Mtbc WGS is now mature 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 standardisation and validation efforts. This will require political commitment, and 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 towards the ethical implications and consequences of routine WGS sequencing and the information it provides. There is thus 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.

Acknowledgements

CJM and LR are also affiliated with BCCM/ITM Mycobacterial Culture Collection, Institute of Tropical Medicine, Antwerp, Belgium; JG is also affiliated with BC Centre for Disease Control, Vancouver, Canada. SG is also affiliated with the University of Basel, Switzerland. BOA is also affiliated with Center for Global Health Security and Diplomacy, Ottawa, Canada. TAK, CU, VD and SN are also affiliated with the German Center for Infection Research, Partner Site Hamburg-Lübeck-Borstel-Riems, D-23845 Borstel, Germany. MS is also affiliated with the University of Arizona, Tuscon, USA. IC is also affiliated to the CIBER in Epidemiology and Public Health, Spain.

Funding:

CJM, BO, LR and BCdJ are supported by an ERC grant [INTERRUPTB; no. 311725]. IC and GAG are supported by an ERC grant (TB-ACCELERATE ; no. 638553).TR receives salary support from the not-for-profit organization Foundation for Innovative New Diagnostics (FIND). The terms of this arrangement have been reviewed and approved by UCSD. TMW is an NIHR Academic Clinical Lecturer. JLG and JG receive funding from the University of British Columbia, Vancouver, Canada. TAK, CU, VD and SN receive funding from the German Center for Infection Research (DZIF) and funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany`s Excellence Strategy – EXC 22167-390884018. LV, THH and AvR are funded by FWO Oddyseus G0F8316N. MF is supported by NIH BD2K K01 (MRF ES026835). PS is supported by the Agence Nationale de la Recherche (ANR-16-CE35-0009), and was a consultant for Genoscreen.

Box 1 : Primary Mtbc diagnostics

Solid or liquid culture (e.g. MGIT, Beckton Dickinson, USA[139]) are the conventional diagnostics for 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 (e.g. pyrazinamide). Therefore, several molecular tests (besides WGS) directly applicable on clinical samples have been developed. Line probe assays rely on hybridization of amplified mycobacterial DNA with nucleotide probes on strips to detect selected drug resistance-associated mutations or their wild-type alleles. MTBDRplus[140,141], TB NTM+MDR[141,142] and MTBDRsl[143,144] were endorsed by WHO. The two former assays target mutations associated with resistance to rifampicin (in *rpoB*) and isoniazid (*katG*, *inhA*), i.e. detect MDR-TB. The MTBDRsl[143,144] assay targets mutations associated with resistance to fluoroquinolones (*gyrA*, *gyrB*) and injectables (*rrs*, *eis*), i.e. detect XDR-TB. Other tests use cartridge-based real-time PCR (GeneXpert MTB-Rif[88,145] (and updated Ultra[146,147]); Anyplex II MDR/XDR[148]; FluoroType MTBDR[149], Hain) or PCR melt-curve (Meltpro[150]) for mutation detection. The FluoroType as well as the WHO-endorsed and globally deployed GeneXpert both detect rifampicin-associated mutations in *rpoB*, plus in the first case, isoniazid resistance mutations (*katG*, *inhA*, *ahpC)*. Because all aforementioned molecular tests use indirect sequencing technologies, they are intrinsically limited to the detection of common pre-selected mutations and are prone to false positive results due to indiscriminate detection of unrelated mutations[151,152]. To circumvent these limitations, newer assays use targeted amplicon sequencing. The Next Gen-RDST[153,154] and Deeplex-MycTB[155,156] assays are directly applicable on clinical samples and sequences (some with promoter regions) of 6 or 18 genes associated with resistance to 7 or 13 anti-tuberculosis drugs, respectively. Deeplex-MycTB additionally includes mycobacterial species and spoligotyping. The large coverage depths that can be achieved enables high confidence mutation calls, including those born by minor subpopulations in case of heteroresistance. Nevertheless, accessible targets are inherently fewer than with WGS.

Glossary terms

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

Spoligotyping: a PCR-based approach based on the amplification of spacers in the CRISPR region of Mycobacterium tuberculosis complex. It is used for genotyping Mtbc strains.

MIRU-VNTR: Mtbc-specific variable tandem repeats loci used to genotype Mtbc strains

cgMLST: core genome multi-locus sequence typing; a scheme that converts genome-wide SNP data into an allele-numbering system using a pre-selected set of core genes

wgMLST: whole genome multi-locus sequence typing; a scheme that converts genome-wide SNP data into an allele-numbering system using a pre-selected set of core genes and additional accessory genes

Löwenstein-Jensen: is a selective culture media in Mycobacteria and commonly used to isolate Mtbc strains

MGIT: the Mycobacteria Growth Indicator Tube is tube that contains mycobacteria selective culture media and which is usually coupled to automated instrument to read the results

Drug susceptibility testing: a procedure to determine if clinical isolates are resistant to antibiotics either by testing the inhibition in culture or by identifying drug resistance associated mutations

SNPs: Single nucleotide polymorphisms; differences in the nucleotide composition of a strain, often compared to a reference (e.g. H37Rv).

WGS workflow: all steps involved (from culturing to SNP calling and analyses) for whole genome sequencing of an isolate

WGS pipeline: the bioinformatics section of the WGS workflow, starting from fastQ files through to SNP calling and analyses

Highlighted references

1. 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. (2018).
2. Starks, A. M. et al. Collaborative Effort for a Centralized Worldwide Tuberculosis Relational Sequencing Data Platform. Clin. Infect. Dis. 61, S141–S146 (2015).
3. The CRyPTIC Consortium and the 100000 Genomes project. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. N. Engl. J. Med. NEJMoa1800474 (2018). doi:10.1056/NEJMoa1800474
4. Coll, F. et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med. 7, 51 (2015).

5. 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).
6. 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).1
7. Coll, F. et al. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. Nat. Commun. 5, 4812 (2014).
8. 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).
9. 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. doi.org 199570 (2017). doi:10.1101/199570
10. Hatherell, H.-A. et al. Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. BMC Med. 14, 21 (2016).

References

1. WHO. *Global tuberculosis report 2018*. (2018).
2. The CRyPTIC Consortium and the 100000 Genomes project. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. *N. Engl. J. Med.* NEJMoa1800474 (2018). doi:10.1056/NEJMoa1800474
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).
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).
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. in *Advances in experimental medicine and biology* **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).
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.* (2018). doi:10.1016/S1473-3099(18)30004-5
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).
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.* (2018). doi:10.1016/S1473-3099(18)30073-2
20. Gröschel, M. I. *et al.* Pathogen-based precision medicine for drug-resistant tuberculosis. *PLOS Pathog.* **14,** e1007297 (2018).
21. Anthony, R., Kamst, M., Nikolayevskyy, V. & van Soolingen, D. *External Quality Assessment of Mycobacterium Interspersed Repetitive Units - Variable Number of Tandem Repeats (MIRU-VNTR) typing and Whole Genome Sequencing analysis of Mycobacterium tuberculosis complex isolates across the European Reference Laboratory* . (2018).
22. 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*. (2018).
23. 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,** 331–42 (2014).
24. Sandgren, A. *et al.* Tuberculosis Drug Resistance Mutation Database. *PLoS Med.* **6,** e1000002 (2009).
25. Coll, F. *et al.* Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med.* **7,** 51 (2015).
26. 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).
27. Starks, A. M. *et al.* Collaborative Effort for a Centralized Worldwide Tuberculosis Relational Sequencing Data Platform. *Clin. Infect. Dis.* **61,** S141–S146 (2015).
28. Brown, T., Nikolayevskyy, V., Velji, P. & Drobniewski, F. Associations between Mycobacterium tuberculosis strains and phenotypes. *Emerg. Infect. Dis.* **16,** 272–80 (2010).
29. Comas, I. *et al.* Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nat. Genet.* **45,** 1176–82 (2013).
30. Meehan, C. J. *et al.* The relationship between transmission time and clustering methods

in Mycobacterium tuberculosis epidemiology. *EBioMedicine* **37,** 410–416 (2018).

31. Kohl, T. A. *et al.* Harmonized Genome Wide Typing of Tubercle Bacilli Using a Web-Based Gene-By-Gene Nomenclature System. *EBioMedicine* **0,** (2018).
32. Walker, T. M. *et al.* Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *Lancet Infect. Dis.* **13,** 137–46 (2013).
33. Koster, K. J. *et al.* Genomic sequencing is required for identification of tuberculosis transmission in Hawaii. *BMC Infect. Dis.* **18,** 608 (2018).
34. Kohl, T. A. *et al.* MTBseq: a comprehensive pipeline for whole genome sequence analysis of *Mycobacterium tuberculosis* complex isolates. *PeerJ* **6,** e5895 (2018).
35. Ezewudo, M. *et al.* Integrating standardized whole genome sequence analysis with a global Mycobacterium tuberculosis antibiotic resistance knowledgebase. *Sci. Rep.* **8,** 15382 (2018).
36. Brynildsrud, O. B. *et al.* Global expansion of *Mycobacterium tuberculosis* lineage 4 shaped by colonial migration and local adaptation. *Sci. Adv.* **4,** eaat5869 (2018).
37. Brown, A. C. *et al.* Rapid Whole-Genome Sequencing of Mycobacterium tuberculosis Isolates Directly from Clinical Samples. *J. Clin. Microbiol.* **53,** 2230–2237 (2015).
38. Conceição, E. C. *et al.* Analysis of potential household transmission events of tuberculosis in the city of Belem, Brazil. *Tuberculosis* **113,** 125–129 (2018).
39. 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).
40. 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. *bioRxiv* 403824 (2018). doi:10.1101/403824
41. Menardo, F. *et al.* Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics* **19,** 164 (2018).
42. Bryant, J. M. *et al.* Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data. *BMC Infect. Dis.* **13,** 110 (2013).
43. 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).
44. 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).
45. 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).
46. 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).
47. 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).
48. Bradley, P. *et al.* Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nat. Commun.* **6,** 10063 (2015).

49. 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 (Edinb).* **95,** 843–844 (2015).

50. 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).

51. Farhat, M. *et al.* genTB: Translational Genomics of Tuberculosis. (2015).

52. 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).

53. Ngo, T.-M. & Teo, Y.-Y. Genomic prediction of tuberculosis drug-resistance: benchmarking existing databases and prediction algorithms. *BMC Bioinformatics* **20,** 68 (2019).

54. 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).

55. 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).

56. 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. *F1000Research* **7,** (2018).

57. FDA. *Infectious Disease Next Generation Sequencing Based Diagnostic Devices: Microbial Identification and Detection of Antimicrobial Resistance and Virulence Markers*. (2016).

58. Pouseele, H. & Supply, P. Accurate Whole-Genome Sequencing-Based Epidemiological Surveillance of Mycobacterium Tuberculosis. *Methods Microbiol.* **42,** 359–394 (2015).

59. 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).

60. Alterovitz, G. *et al.* Enabling precision medicine via standard communication of HTS provenance, analysis, and results. *PLOS Biol.* **16,** e3000099 (2018).

61. Stucki, D. *et al.* Standard Genotyping Overestimates Transmission of Mycobacterium tuberculosis among Immigrants in a Low-Incidence Country. *J. Clin. Microbiol.* **54,** 1862–70 (2016).

62. 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).

63. 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).

64. Coll, F. *et al.* Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. *Nat. Genet.* **50,** 307–316 (2018).

65. Farhat, M. R. *et al.* Genome wide association with quantitative resistance phenotypes in Mycobacterium tuberculosis reveals novel resistance genes and regulatory regions. *Nat. Commun.* (2019). doi:10.1101/429159

66. Kwong, J. C., Mccallum, N., Sintchenko, V. & Howden, B. P. Whole genome sequencing in clinical and public health microbiology. *Pathology* **47,** 199–210 (2015).

67. 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. *doi.org* 199570 (2017). doi:10.1101/199570

68. 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.* (2019). doi:10.1093/cid/ciz219

69. 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,** S27 (2010).

70. 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).

71. World Health Organization. *WHO's code of conduct for open and timely sharing of pathogen genetic sequence data during outbreaks of infectious disease*. (2019).

72. 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).

73. Karikari, T. K. Bioinformatics in Africa: The Rise of Ghana? *PLoS Comput. Biol.* **11,** e1004308 (2015).

74. Tekola-Ayele, F. & Rotimi, C. N. Translational Genomics in Low- and Middle-Income Countries: Opportunities and Challenges. *Public Health Genomics* **18,** 242–247 (2015).

75. Helmy, M., Awad, M. & Mosa, K. A. Limited resources of genome sequencing in developing countries: Challenges and solutions. *Appl. Transl. Genomics* **9,** 15–19 (2016).

76. 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).

77. Bradley, P. *et al.* Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nat. Commun.* **6,** 10063 (2015).

78. Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: Scientific containers for mobility of compute. *PLoS One* **12,** e0177459 (2017).

79. Merkel, D. Docker: lightweight linux containers for consistent development and deployment. *Linux J.* **2014,** 2 (2014).

80. Grüning, B. *et al.* Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* **15,** 475–476 (2018).

81. Jackman, S., Birol, I., Jackman, S. & Birol, I. Linuxbrew and Homebrew for cross-platform package management. *F1000Research* **5,** (2016).

82. Langille, M. G. I. & Eisen, J. A. BioTorrents: a file sharing service for scientific data. *PLoS One* **5,** e10071 (2010).

83. Karikari, T. K., Quansah, E. & Mohamed, W. M. Y. Widening participation would be key in enhancing bioinformatics and genomics research in Africa. *Appl. Transl. genomics* **6,** 35–41 (2015).

84. 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).

85. Zignol, M. *et al.* Population-based resistance of Mycobacterium tuberculosis isolates to pyrazinamide and fluoroquinolones: results from a multicountry surveillance project. *Lancet Infect. Dis.* (2016). doi:10.1016/S1473-3099(16)30190-6

86. Kumwenda, S. *et al.* Challenges facing young African scientists in their research careers: A qualitative exploratory study. *Malawi Med. J.* **29,** 1–4 (2017).

87. 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).

88. Helb, D. *et al.* Rapid detection of Mycobacterium tuberculosis and rifampin resistance by use of on-demand, near-patient technology. *J. Clin. Microbiol.* **48,** 229–37 (2010).

89. 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).

90. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15,** R46 (2014).

91. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26,** 1721–1729 (2016).

92. 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).

93. Periwal, V. *et al.* Comparative whole-genome analysis of clinical isolates reveals characteristic architecture of Mycobacterium tuberculosis pangenome. *PLoS One* **10,** e0122979 (2015).

94. Gao, Q. *et al.* Gene expression diversity among Mycobacterium tuberculosis clinical isolates. *Microbiology* **151,** 5–14 (2005).

95. Kato-Maeda, M. *et al.* Comparing genomes within the species Mycobacterium tuberculosis. *Genome Res.* **11,** 547–54 (2001).

96. 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).

97. Ioerger, T. R. *et al.* Variation among genome sequences of H37Rv strains of Mycobacterium tuberculosis from multiple laboratories. *J. Bacteriol.* **192,** 3645–53 (2010).

98. Lee, R. S. & Behr, M. A. Does Choice Matter? Reference-Based Alignment for Molecular Epidemiology of Tuberculosis. *J. Clin. Microbiol.* **54,** 1891–1895 (2016).

99. Norman, A., Folkvardsen, D. B., Overballe-Petersen, S. & Lillebaek, T. Complete genome sequence of Mycobacterium tuberculosis DKC2, the predominant Danish outbreak strain. *Genome Announc.* **8,** e01554-18 (2019).

100. 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).

101. 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. Genomics* **4,** 256719 (2018).

102. 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).

103. Yadon, A. N. *et al.* A comprehensive characterization of PncA polymorphisms that confer resistance to pyrazinamide. *Nat. Commun.* **8,** 588 (2017).

104. Yang, Y. *et al.* Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. *Bioinformatics* **34,** 1666–1671 (2018).

105. Chen, M. L. *et al.* Deep learning predicts tuberculosis drug resistance status from genome sequencing data. *bioRxiv* 275628 (2018). doi:10.1101/275628

106. Rajendran, V. & Sethumadhavan, R. Drug resistance mechanism of PncA in Mycobacterium tuberculosis. *J. Biomol. Struct. Dyn.* (2013).

107. 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).

108. Duchêne, S. *et al.* Genome-scale rates of evolutionary change in bacteria. *Microb. genomics* **2,** e000094 (2016).

109. Lee, R. S. *et al.* Reemergence and Amplification of Tuberculosis in the Canadian Arctic. *J. Infect. Dis.* **211,** 1905–1914 (2015).

110. 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).

111. 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).

112. 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,** 786–92 (2013).

113. 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).

114. 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).

115. Lieberman, T. D. *et al.* Genomic diversity in autopsy samples reveals within-host dissemination of HIV-associated Mycobacterium tuberculosis. *Nat. Med.* (2016). doi:10.1038/nm.4205

116. 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,** 784–90 (2013).

117. Ford, C. B. *et al.* Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. *Nat. Genet.* **43,** 482–6 (2011).

118. Hatherell, H.-A. *et al.* Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. *BMC Med.* **14,** 21 (2016).

119. Verver, S. *et al.* Transmission of tuberculosis in a high incidence urban community in

South Africa. *Int. J. Epidemiol.* **33,** 351–357 (2004).

120. 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).

121. Stimson, J. *et al.* Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions. *Mol. Biol. Evol.* **36,** 587–603 (2019).

122. 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).

123. Campbell, F. *et al.* outbreaker2: a modular platform for outbreak reconstruction. *BMC Bioinformatics* **19,** 363 (2018).

124. Didelot, X., Gardy, J. & Colijn, C. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol. Biol. Evol.* **31,** 1869–79 (2014).

125. Didelot, X., Fraser, C., Gardy, J. & Colijn, C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Mol. Biol. Evol.* **34,** msw075 (2017).

126. 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).

127. 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).

128. Kühnert, D. *et al.* Tuberculosis outbreak investigation using phylodynamic analysis. *Epidemics* **25,** 47–53 (2018).

129. Eldholm, V. *et al.* Armed conflict and population displacement as drivers of the evolution and dispersal of Mycobacterium tuberculosis. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 13881–13886 (2016).

130. Streicher, E. M. *et al.* Mycobacterium tuberculosis population structure determines the outcome of genetics-based second-line drug resistance testing. *Antimicrob. Agents Chemother.* **56,** 2420–7 (2012).

131. 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,** 4220–2 (2013).

132. 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).

133. Sobkowiak, B. *et al.* Identifying mixed Mycobacterium tuberculosis infections from whole genome sequence data. *BMC Genomics* **19,** 613 (2018).

134. 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).

135. 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).

136. 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).

137. 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).

138. Phelan, J. E. *et al.* Recombination in pe/ppe genes contributes to genetic variation in Mycobacterium tuberculosis lineages. *BMC Genomics* **17,** 151 (2016).

139. 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,** 325–9 (1995).

140. 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,** 3032–4 (2015).

141. 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).

142. 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).

143. 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,** 1767–72 (2009).

144. 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).

145. 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. informatics* **6,** e12 (2018).

146. 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).

147. Ng, K. C. S. *et al.* Xpert Ultra Can Unambiguously Identify Specific Rifampin Resistance-Conferring Mutations. *J. Clin. Microbiol.* **56,** e00686-18 (2018).

148. Molina-Moya, B. *et al.* Diagnostic accuracy study of multiplex PCR for detecting tuberculosis drug resistance. *J. Infect.* **71,** 220–230 (2015).

149. 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).

150. Pang, Y. *et al.* Rapid diagnosis of MDR and XDR tuberculosis with the MeltPro TB assay in China. *Sci. Rep.* **6,** 25330 (2016).

151. 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,** 2876–80 (2014).

152. 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).

153. 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).

154. 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).

155. 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).

156. 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).

Display item legends

Figure 1: The primary tasks for whole genome sequencing in public health. Assessing the epidemiology (surveillance and clustering/outbreaks) and determining the strain type or resistance profile to specific drugs can all be undertaken using the genomic variant calls derived from Mtbc WGS pipelines.

Figure 2: Common workflow for whole genome sequencing for Mtbc isolates. A clinical sample (often sputum) is first cultured for up to 6 weeks followed by gDNA extraction and sequencing. The resulting sequencing output (fastq files) can be deposited online to public repositories and also run through standard SNP-calling pipelines which will undertake read mapping and variant calling. The resulting SNP lists can then be used for a variety of analyses, each which then can be reported to the end user.

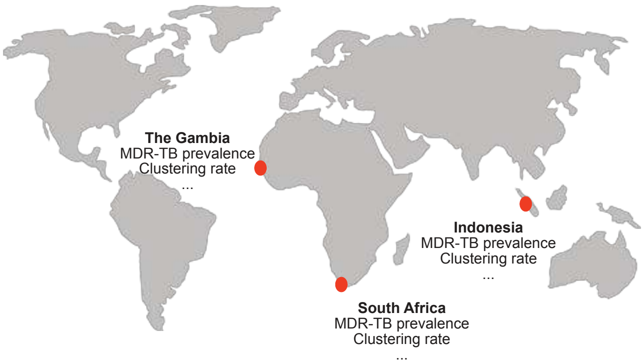
Figure 3: Current and potential future approach for determining resistance-related polymorphisms. In the current approach (green box), lists of resistance-related SNPs are primarily built using a statistical approach, often a likelihood ratio. This uses linked phenotypic/genotypic data derived from a variety of strains across the diversity of the Mtbc to create lists of known SNPs that cause drug resistance. The suggested extension (blue box) would complement this procedure with additional information from targeted mutagenesis etc. to detect drug resistance causing SNPs too rare to be detected using a statistical approach.

Figure 4: Epidemiological and within-host applications of SNP-based comparisons between Mtbc isolates. At a population level, SNP-based phylogenetics can be used to recreate local diversity. These phylogenies are then sub-divided into transmission clusters using pre-defined SNP or allele cut-offs. At the individual level, within-host diversity can be generated either through sub-population divergence or infection with multiple concurrent strains.

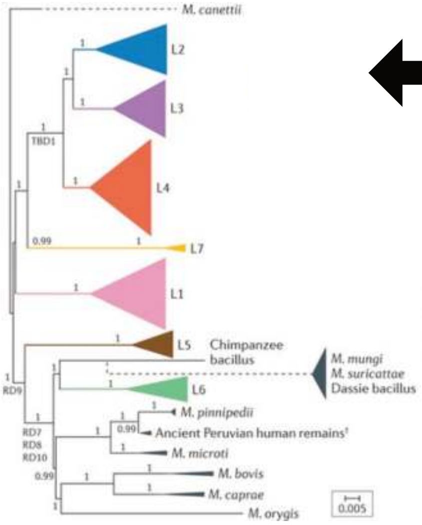
Supplementary table 1: A non-exhaustive list of common bioinformatics pipelines and their settings for SNP calling of Mtbc isolates. This list contains only a small portion of the available pipelines but demonstrates the variability and breadth of the field.

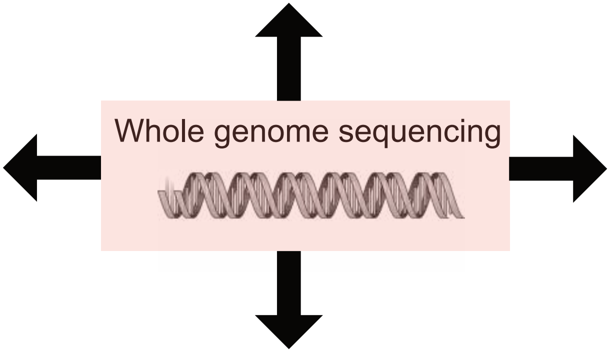
Supplementary table 2: Suggested elements and attributes for standardised reporting of Mtbc WGS result

Surveillance



Strain and subspecies typing

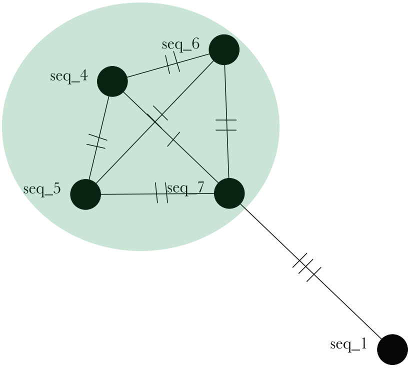


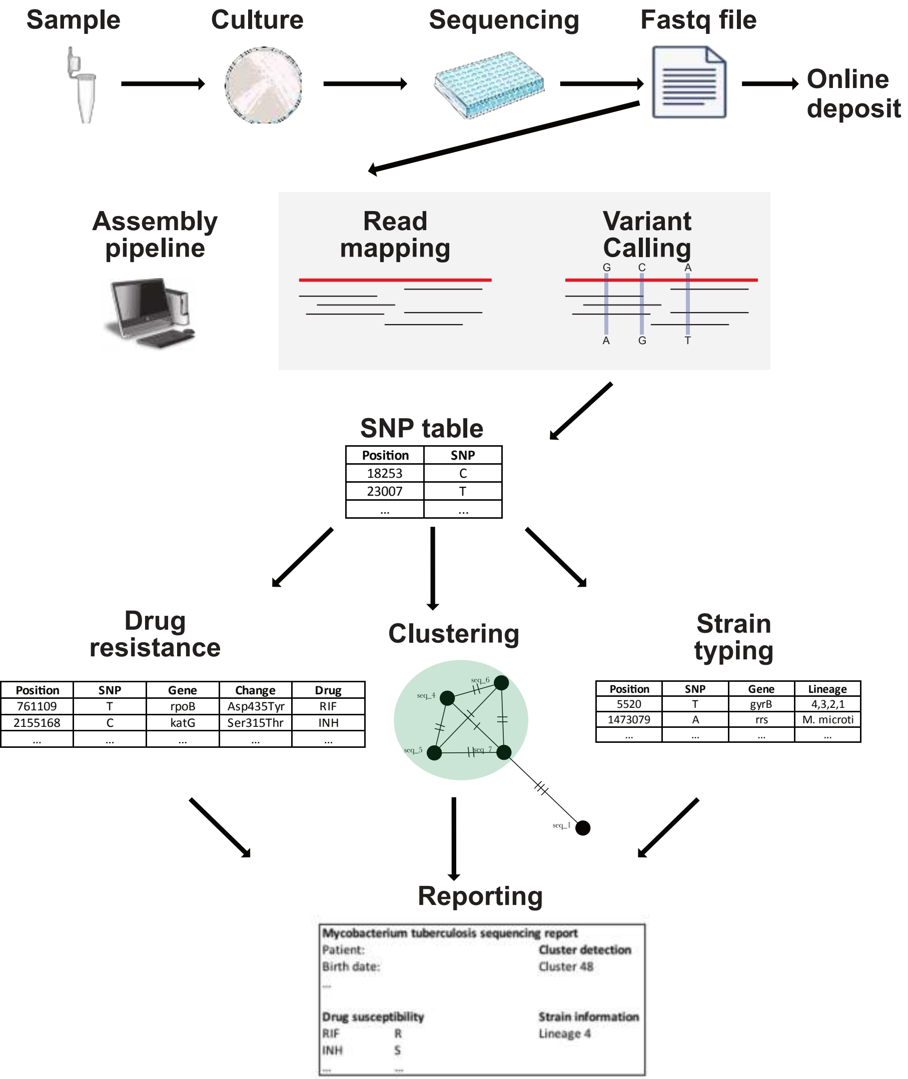


Drug susceptibility

Drug	Resistance
RIF	R
INH	R
ETH	S
PZA	S

Clustering and outbreaks





Position	SNP
18253	C
23007	T
...	...

Position	SNP	Gene	Change	Drug
761109	T	rpoB	Asp435Tyr	RIF
2155168	C	katG	Ser315Thr	INH
...	...	...	...	...

Position	SNP	Gene	Lineage
5520	T	gyrB	4,3,2,1
1473079	A	rrs	M. microti
...	...	...	...

Mycobacterium tuberculosis sequencing report

Patient:

Birth date:

...

Cluster detection

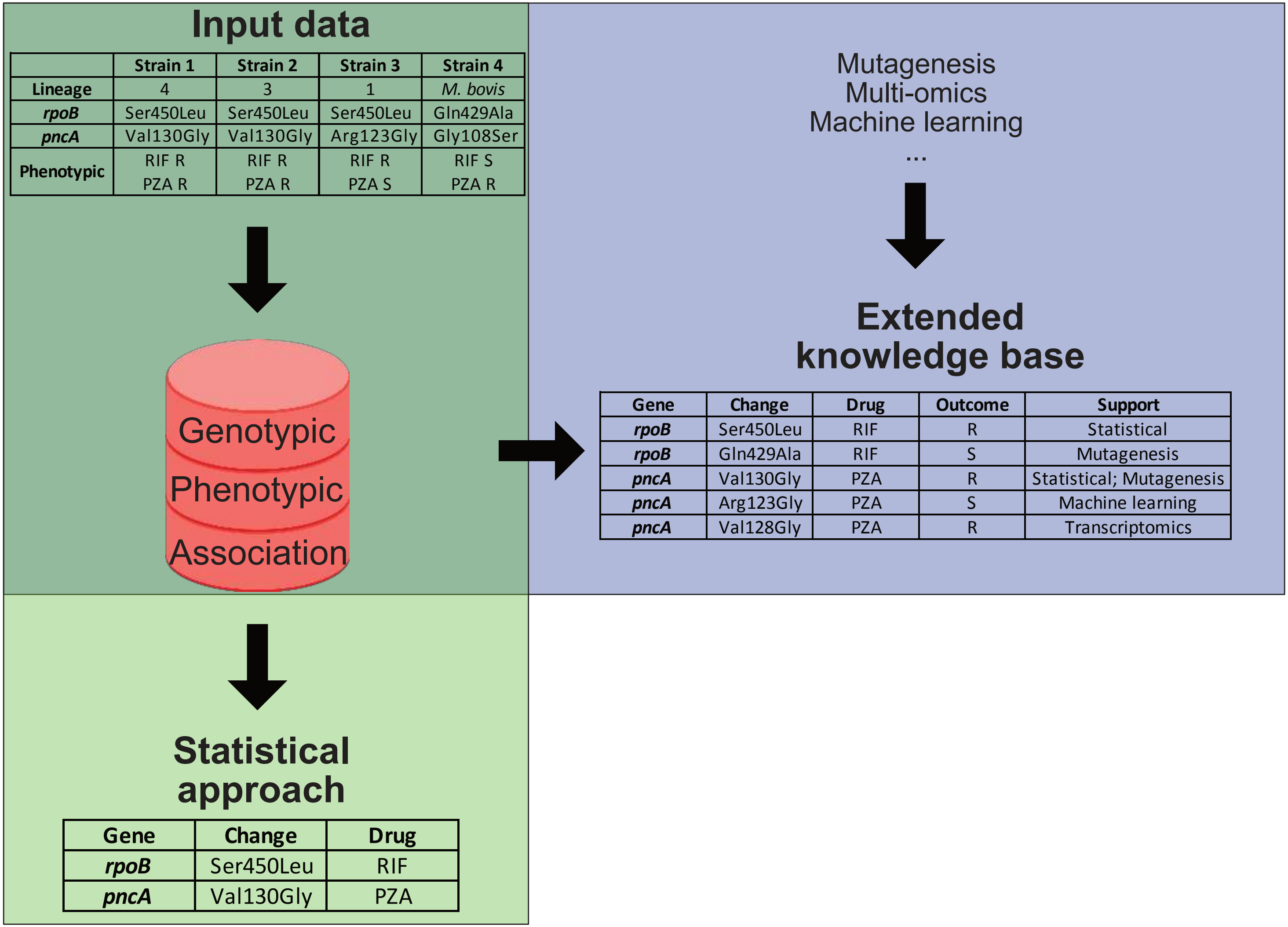
Cluster 48

Drug susceptibility

RIF	R
INH	S
...	...

Strain information

Lineage 4



Input data

	Strain 1	Strain 2	Strain 3	Strain 4
Lineage	4	3	1	*M. bovis*
rpoB	Ser450Leu	Ser450Leu	Ser450Leu	Gln429Ala
pncA	Val130Gly	Val130Gly	Arg123Gly	Gly108Ser
Phenotypic	RIF R PZA R	RIF R PZA R	RIF R PZA S	RIF S PZA R

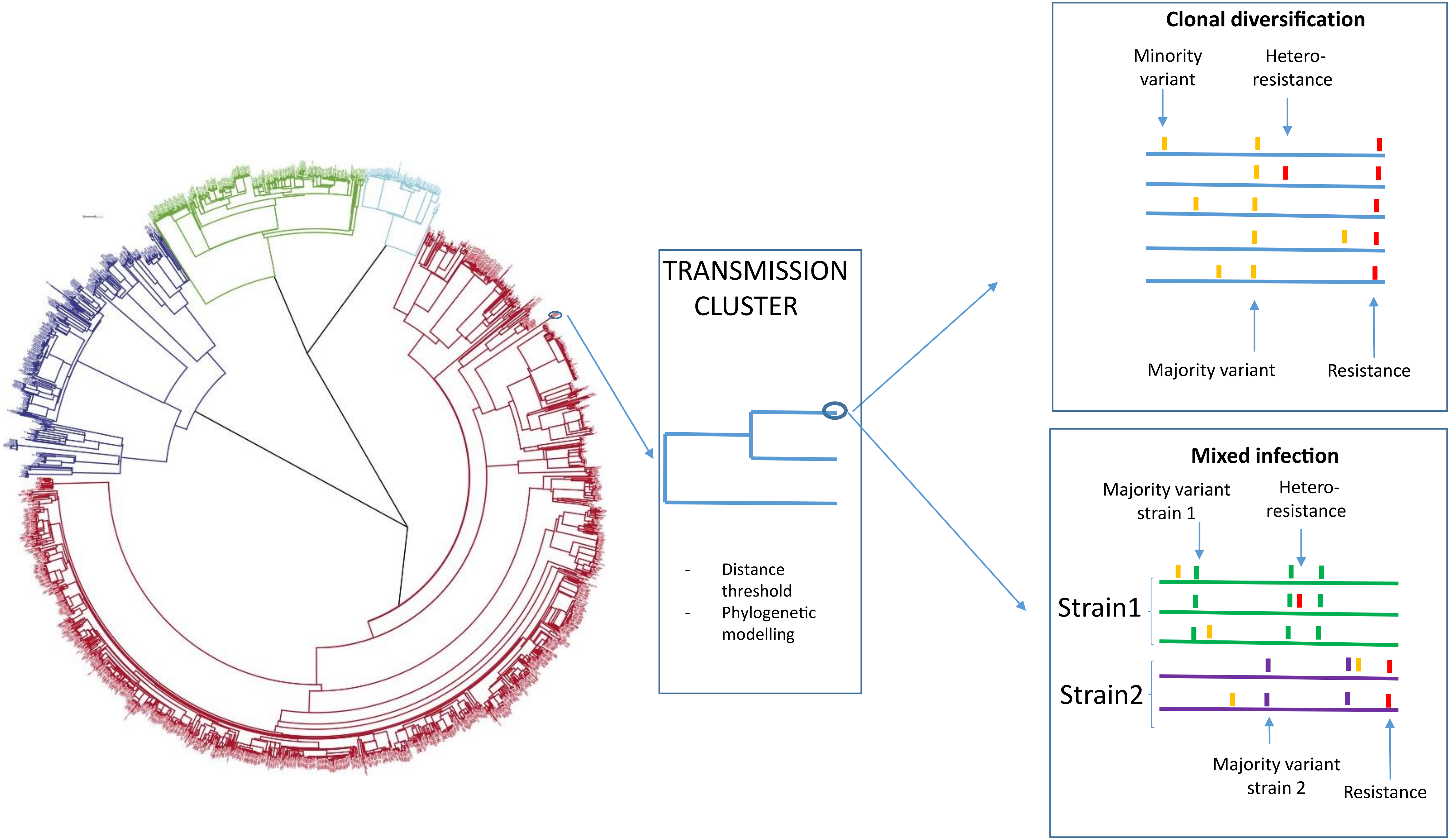
Statistical approach

Gene	Change	Drug
rpoB	Ser450Leu	RIF
pncA	Val130Gly	PZA

Mutagenesis
Multi-omics
Machine learning
...

Extended knowledge base

Gene	Change	Drug	Outcome	Support
rpoB	Ser450Leu	RIF	R	Statistical
rpoB	Gln429Ala	RIF	S	Mutagenesis
pncA	Val130Gly	PZA	R	Statistical; Mutagenesis
pncA	Arg123Gly	PZA	S	Machine learning
pncA	Val128Gly	PZA	R	Transcriptomics



TRANSMISSION CLUSTER

- Distance threshold
- Phylogenetic modelling

Clonal diversification

Minority variant

Hetero-resistance

Majority variant

Resistance

Mixed infection

Majority variant strain 1

Hetero-resistance

Strain1

Strain2

Majority variant strain 2

Resistance