Evaluating 12 automated, whole-genome sequencing analysis pipelines for *Mycobacterium tuberculosis* complex: a comparative study

Ruan Spies, Derrick W Crook, Timothy E A Peto, Philip W Fowler, Robert Turner, Hieu Thai, James A Watson, Timothy M Walker

Summary

Background Reliance on complex, custom-built bioinformatics pipelines is a barrier to the implementation of whole-genome sequencing (WGS) of *Mycobacterium tuberculosis* in high-burden settings in some low-income and middle-income countries (LMICs). Automated analysis pipelines could address this inequity in access to WGS-based diagnostics and surveillance. This study aimed to systematically evaluate the performance and usability of publicly available WGS pipelines for *M tuberculosis*.

Methods We identified automated *M tuberculosis* WGS analysis pipelines through searches of PubMed and GitHub from database inception up to Aug 31, 2024. Accuracy, cost, accessibility, and scalability were assessed for each pipeline. We evaluated the accuracy of genotypic drug susceptibility testing (gDST) using publicly available sequences with phenotypic susceptibility data for 12 antituberculosis drugs. We estimated pooled sensitivity and specificity for each pipeline, across all drugs, by conducting a bivariate meta-analysis, with random effects representing between-drug variability. Lineage classifications were compared, and a previously epidemiologically well-characterised dataset was used to compare measures of genomic relatedness.

Findings Among 28 candidate pipelines, 16 were excluded as they were unmaintained and inexecutable. 12 pipelines (11 compatible with Illumina and four compatible with Nanopore), all free to use, were included for evaluation. Six pipelines processed and stored data remotely, but for five of these six, scalability was limited by the need to upload sequences through web portals. For local processing pipelines, scalability was dependent on substantial local computational resources, data storage capacity, and command-line interfaces that limited user-friendliness. Only one of six remote-processing pipelines removed human DNA sequences before server upload. gDST was similarly accurate across ten of 11 Illumina-compatible pipelines and three of four Nanopore-compatible pipelines. All pipelines classified the main lineages consistently, although there were differences at sublineage resolution. Outputs from three of four pipelines reporting genomic relatedness were compatible with commonly cited single nucleotide polymorphism difference thresholds.

Interpretation Numerous automated analysis pipelines capable of enhancing equity in *M tuberculosis* WGS are available. Given the overall similarities between the pipelines evaluated in this study in terms of gDST performance, lineage classification, and genomic relatedness inference, non-functional attributes such as availability, accessibility, scalability, and privacy could represent the point of difference for prospective users in LMICs with a high burden of tuberculosis.

Funding The Rhodes Trust, Wellcome, Ellison Institute of Technology, and the UK National Institute for Health and Care Research Oxford Biomedical Research Centre.

Lancet Microbe 2025; 6: 101210

Published Online October 21, 2025
https://doi.org/10.1016/j.lanmic.2025.101210

Oxford University Clinical Research Unit, Ho Chi Minh City, Viet Nam (R Spies MSc, T M Walker DPhil); **Nuffield Department of Medicine, University of Oxford, Oxford, UK** (R Spies, Prof D W Crook FRCPath, Prof T E A Peto FRCP, P W Fowler PhD, R Turner PhD, H Thai BE, J A Watson DPhil, T M Walker); **National Institute of Health Research Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, UK** (Prof D W Crook, P W Fowler); **Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, Oxford, UK** (Prof D W Crook, P W Fowler)

Correspondence to:
Dr Ruan Spies, Oxford University Clinical Research Unit, Ho Chi Minh City 700000, Viet Nam
rspies@oucru.org

Introduction

The first routine whole-genome sequencing (WGS) service for *Mycobacterium tuberculosis* was established in the UK in 2017.[1] Similar services in other high-income countries followed as the potential of WGS for tuberculosis control became clear.[2] However, each service has had to build its own infrastructure to process clinical samples through the laboratory and return reports to clinicians in a timely manner. In each case, this process has involved building complex, in-house bioinformatics pipelines de novo. However, most patients with tuberculosis do not live in countries with this capacity, further exacerbating global inequity in access to diagnostic and surveillance tools.[3]

One response to the COVID-19 pandemic was to increase sequencing capacity in a number of low-income and middle-income countries (LMICs).[4] There is potential to harness this legacy by repurposing technology towards the older but more intransigent tuberculosis pandemic, but hurdles such as data processing and interpretation need to be overcome by administrations seeking to build WGS services.[5,6] Off-the-shelf options now exist because different academic groups have generated automated *M tuberculosis*

Research in context

Evidence before this study

We searched MEDLINE and Google Scholar for any publications evaluating automated *Mycobacterium tuberculosis* complex whole-genome sequencing (WGS) analysis tools. We searched for articles published between Jan 1, 2013, and Aug 31, 2024, with no language restrictions. We identified five studies that quantitatively compared resistance prediction between pipelines and two studies that were narrative reviews of bioinformatics tools applicable to *M tuberculosis* WGS. Previous studies generally had small sample sizes, only included a few established pipelines, and only included Illumina-sequenced isolates. Indeed, among six novel pipelines released since 2022, only one (SAM-TB) was evaluated, by a single study. Studies have generally demonstrated excellent accuracy across pipelines in WGS-based resistance prediction for rifampicin and isoniazid, with lower accuracy and more variability observed for pyrazinamide, ethambutol, second-line injectable drugs, and fluoroquinolones. No previous studies have evaluated WGS-based resistance prediction for new and repurposed drugs such as bedaquiline and linezolid. Only one study evaluated lineage classification in addition to resistance prediction, observing differences between pipelines and attributing the differences to the resolution to which lineage was classified. A single study compared genomic relatedness inferred between five in-house pipelines, but none of these were publicly available, automated pipelines.

Added value of this study

We applied a systematic search strategy using PubMed and GitHub to identify published and unpublished pipelines, and coauthors with subject expertise proposed additional pipelines not identified by the initial search. As a result, we included and evaluated twelve pipelines—more than double the number evaluated by previous studies. The scope of our evaluation exceeded previous studies, as we not only assessed pipelines according to how well they predicted resistance to different antituberculosis drugs but also evaluated their accuracy when assigning lineage and identifying genomically related samples that could be part of the same epidemiological cluster. To our knowledge, this is the first study to have identified and evaluated long read-compatible (Oxford Nanopore Technologies) pipelines. In addition to this quantitative measure of accuracy, we describe and evaluate other important pipeline attributes such as availability, accessibility, scalability, and privacy considerations.

Implications of all the available evidence

Numerous free-to-use, automated analysis pipelines that might improve equity in access to *M tuberculosis* WGS are currently available. Although the accuracy of WGS-based drug resistance prediction, lineage classification, and genomic relatedness is broadly similar for most pipelines, there is greater variation in user-friendliness, scalability, and other non-functional requirements between pipelines. Availability, accessibility, scalability, privacy, security, and sustainability, therefore, are key considerations that could determine which option the tuberculosis programmes in high-burden settings in some low-income and middle-income countries might wish to consider when selecting a free-to-use automated *M tuberculosis* analysis pipeline.

analysis pipelines. However, for potential implementers to make informed decisions around the optimal tool for their setting, a thorough evaluation of available options is required.

Here, we present a systematic evaluation and validation of 12 automated *M tuberculosis* bioinformatics pipelines, with a focus on drug susceptibility test predictions, species and lineage identification, identification of genomically related samples, availability, accessibility, scalability, and privacy considerations.

Methods

Pipeline eligibility and search and selection process

We defined automated *M tuberculosis* WGS analysis pipelines as those that could be deployed without the manual execution of multiple workflow stages. At a minimum, eligible pipelines were required to accept raw sequencing reads (FASTQ files) as inputs and to conduct genotypic drug susceptibility testing (gDST). To identify existing pipelines, RS searched PubMed and GitHub from database inception to Aug 31, 2024, using the search terms shown in appendix 1 (p 4). The study also included two pipelines that were not identified through the search but were already known to the authors. Pipeline developers were contacted when troubleshooting was required, with non-response resulting in exclusion of the pipeline from the study. Preliminary insights from this study might have contributed to updates to some pipelines. The final evaluations presented here represent the publicly available, referenced versions of the included pipelines (appendix 1 p 7). This study used publicly available data and did not require ethics approval.

See Online for appendix 1

Pipeline software attributes

We proposed a core set of attributes to be considered by national tuberculosis programmes in LMICs. First, pipelines should be accurate for the most relevant applications of tuberculosis WGS: resistance prediction, species and lineage classification, and genomic relatedness inference. Second, practical considerations were taken into account; these included the cost of the pipeline, how the pipeline is accessed (through a graphical user interface [GUI] and/or a less user-friendly command-line interface [CLI]), where data are stored, how the software is licensed, how human genetic material is handled, when the software was last updated, and whether the pipeline has received any formal accreditation. We assessed each pipeline against these criteria if the data or relevant documentation allowed. Data

security was not considered, as the technical requirements for this aspect vary considerably by application.

Accuracy and functionality

We curated a test dataset comprising 1000 Illumina-sequenced *M tuberculosis* isolates to evaluate resistance prediction for the Illumina-compatible pipelines that permitted bulk sequence processing; the number of isolates was selected as a trade-off between scale and practicality in executing multiple computationally intensive processes. These pipelines were the CLI versions of the Global Pathogen Analysis Service (GPAS), MAGMA, MTBseq, Mykrobe, TBProfiler, TBSeqPipe, and tbtAMR. Metadata were obtained from the CRyPTIC public database,[7] in which all phenotypic drug susceptibility test (pDST) results for 13 antituberculosis drugs were available for all samples, and FASTQ files were downloaded from the European Nucleotide Archive. The dataset was restricted to isolates with high confidence minimum inhibitory concentration (MIC) measurements and enriched for drug-resistant isolates (appendix 1 p 4).

All pipelines were executed using default settings, and the time taken to process all 1000 sequences was recorded (appendix 1 p 5). Rates of false negatives and false positives were calculated for the gDST predictions of each pipeline for rifampicin, isoniazid, ethambutol, moxifloxacin, levofloxacin, amikacin, kanamycin, ethionamide, bedaquiline, linezolid, clofazimine, and delamanid, where these were reported (appendix 1 p 9). The CRyPTIC dataset lacked pDST results for pyrazinamide.[7] The CRyPTIC binary phenotype, inferred from measured MICs, was used as the reference standard, and 95% CIs were calculated using Wilson score intervals with the binom package in R.[8] We estimated pooled sensitivity and specificity for each pipeline, across all drugs, by conducting a bivariate meta-analysis with random effects representing between-drug variability, implemented with the mada package.[9] The meta-analysis excluded bedaquiline and clofazimine, as Mykrobe does not report these drugs.

As not all pipelines fulfilling the study inclusion criteria were compatible with bulk sequence processing, we curated another dataset consisting of 100 sequences from CRyPTIC (appendix 1 p 4).[7] These sequences were processed with all Illumina-compatible pipelines, which additionally included web applications for Gen-TB, GenoMycAnalyzer, PhyResSE, and SAM-TB.

We were unable to identify any publicly available databases as large or as rigorously phenotyped as CRyPTIC for Nanopore-sequenced *M tuberculosis* isolates. Our test dataset for the Nanopore-compatible pipelines was, therefore, limited to 90 isolates. These isolates were sequenced with both Illumina and Nanopore technologies and had pDST results for rifampicin, isoniazid, ethambutol, and streptomycin. The sequence accession numbers and pDST results were obtained from a dataset previously compiled by Hall and colleagues (appendix 1 p 5).[10]

pDST and gDST discordance analyses were done for each dataset. Failures were defined as instances when no gDST prediction was made by a pipeline. False positives were defined as instances when the gDST reported resistance but the pDST reported susceptibility. False negatives were defined as instances when the gDST reported susceptibility and the pDST reported resistance. In cases where pipelines reported an uncertain or unknown result due to an uncharacterised or unclassified mutation in a relevant gene, the prediction was classified as susceptible, to allow comparability across all pipelines. We examined output variant call format files and resistance mutation catalogues to adjudicate the sources of false positive and false negative predictions for pipelines included in the 1000 Illumina sequences evaluation (appendix 1 pp 5–6).

Species identification and lineage classification

We screened the specificity of species identification for *M tuberculosis* among all Illumina-compatible pipelines by uploading paired FASTQ files for ten non-tuberculous mycobacteria as inputs (appendix 1 p 6). Given the absence of a standardised reference for lineage classification, we assessed agreement between pipelines by comparing the classification of isolates into main lineages and sublineages with the same datasets as used for resistance prediction. Agreement between pipelines was quantified with the irr package through the calculation of Fleiss' kappa, which measures the consistency of categorical assignments across multiple raters (appendix 1 p 7).[11]

Genomic relatedness

To evaluate genomic relatedness inference, we used outputs from four Illumina-compatible pipelines to reconstruct genomic distances for a previously published dataset that established what have since become conventional cluster-defining single nucleotide polymorphism (SNP) difference thresholds.[12] This dataset consisted of 30 pairs of longitudinally sampled patient isolates and 38 pairs of household-derived or family-derived isolates, all from a low-prevalence setting in the UK.[12]

MTBseq, SAM-TB, and TBSeqPipe produced multiple sequence alignments of variant sites, whereas GPAS produced FASTA files for individual isolates aligned to the H37Rv (NC_000962.3) reference genome. We concatenated the FASTA files of individual isolates produced by GPAS into a multiple sequence alignment consisting solely of variant sites. We derived pairwise SNP distances from each pipeline's multiple sequence alignment (appendix 1 p 7). Although MAGMA reports on genomic relatedness, producing multiple sequence alignments of variant sites, at the time of writing this pipeline was incompatible with the specific sequences used in this analysis and was excluded.

Role of the funding source

The Ellison Institute of Technology ([EIT] Oxford, UK), the sponsor of the GPAS pipeline, provided access to GPAS for

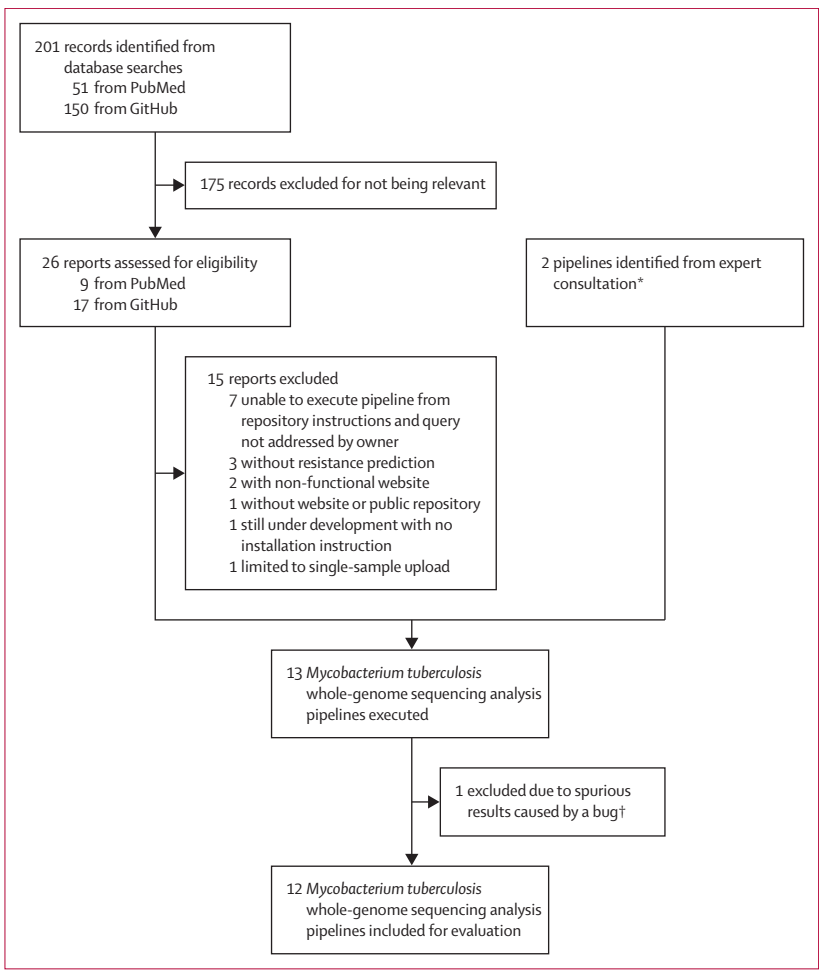

Figure 1: **Pipeline selection**
*Consultation with experts in *Mycobacterium tuberculosis* whole-genome sequencing. †Bug not resolved at the time of writing.

evaluation in this study, before its public release. EIT, through its relationship with Oracle, also facilitated access to the computational infrastructure used to conduct the analyses free of charge. The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

We identified 28 automated *M tuberculosis* analysis pipelines, of which 16 were excluded (for reasons including being unmaintained or inexecutable at the time of writing; figure 1; appendix 1 p 10). Eleven of the included 12 pipelines were compatible with Illumina, four with Nanopore, two with Ion Torrent, and one with PacBio sequence data.

The characteristics of the included pipelines are summarised in the table.[13–23] Eleven pipelines were available free of charge to all users, whereas GPAS was only free for users in LMICs and was available to users in other countries for a trial limited to 100 samples. Three pipelines were accessible through both GUIs and CLIs, four through GUIs only, and another five were accessible through CLIs only. Among the six remote-processing pipelines, only GPAS removed identifiable human reads before server upload. Licences, where specified, were MIT, GPL or AGPL, all of which are open source licences. The MIT licence is the most permissive, whereas GPL requires derivative works to carry the same licence. GPL software can generally be used in the Software as a Service (SaaS) platform without restriction, whereas AGPL denies this option, making SaaS operators share their source code when using AGPL software. Ten pipelines were produced by academic groups. Exceptions were GPAS, a UK academic–private collaboration; tbtAMR, which was co-produced by a national public health laboratory in Australia; and TBSeqPipe, which had unclear origins. Only tbtAMR reported any formal accreditation.[23]

Pipeline functionalities are summarised in appendix 1 (pp 11–12). Although there was variation in the lineage classification schemas and resistance prediction catalogues used, the Coll lineage classification schema[24] was used by seven pipelines, either directly or by those applying TBProfiler for lineage classification. Moreover, versions of WHO resistance mutation catalogues[25,26] were used by ten pipelines, with most applying additional custom interpretations. Seven pipelines reported genomic relatedness, and four generated phylogenetic trees. Additional features of the included pipelines are summarised in appendix 1 (p 13).

gDST accuracies with the 1000 Illumina sequences dataset were similar for the seven CLI-based, Illumina-compatible pipelines (GPAS, MAGMA, MTBseq, Mykrobe, TBProfiler, TBSeqPipe, and tbtAMR) except for MTBseq, which had higher false negative rates for rifampicin, isoniazid, moxifloxacin, levofloxacin, and ethionamide (figure 2; appendix 1 p 14). All pipelines had high false negative rates for the new and repurposed drugs bedaquiline, clofazimine, delamanid, and linezolid, ranging from 42·9% (95% CI 21·4–67·4) for delamanid (MAGMA) to 100% for bedaquiline and clofazimine (TBSeqPipe; 100% [67·6–100·0] for bedaquiline and 100% [90·6–100·0] for clofazimine; appendix 1 p 14). False positive rates for bedaquiline, clofazimine, delamanid, and linezolid were 1% or less for all pipelines. pDST results are presented in appendix 1 (p 15). Pooled sensitivity for all drugs, excluding bedaquiline and clofazimine, ranged from 82·9% (95% CI 67·2–91·9) for MTBseq to 91·0% (83·0–95·5) for MAGMA, with pooled specificity no lower than 96·8% for all pipelines (appendix 1 p 15).

Among the CLI-based Illumina-compatible pipelines, TBProfiler had the lowest overall proportion of predictions that were errors (failures, false negatives, or false positives), at 3·6% (438 errors per 12 000 predictions), and MAGMA had the highest overall proportion, at 6·4% (770 per 12 000 predictions), driven largely by its marginally higher failure rate of 2·6% (312 failures per 12 000 predictions; failure rates for these seven pipelines ranged from 0% for TBProfiler and Mykrobe to 2·6% for MAGMA; appendix 1 p 20).

	Website	Cost	GUI	CLI	Data processing and storage	Software licence	Version and date accessed	Accreditation
GPAS[13]	https://app.gpas.global	Free for users in LMICs; 100-sample free trial for other users	Yes, web portal	Yes	Remote	GPL 3.0 (CLI) and end-user licence agreement	Version 2.0.0; September, 2024	None specified
Gen-TB[14]	https://gentb.hms.harvard.edu/	Free	Yes, web portal	No	Remote	APGL 3·0	Not specified; August, 2024	None specified
GenoMycAnalyzer[15]	https://mycochase.org/	Free	Yes, web portal	No	Remote	GPL 3.0	April, 2024	None specified
MAGMA[16]	https://github.com/TORCH-Consortium/MAGMA	Free	No	Yes	Local	GPL 3.0	Version 2.0.0-alpha; November, 2024	None specified
MTBseq[17]	https://github.com/ngs-fzb/MTBseq_source	Free	No	Yes	Local	GPL 3.0	Version 1.10; August, 2024	None specified
Mykrobe[18]	https://github.com/mykrobe-tools/mykrobe	Free	Yes, offline application; single-sample processing only (no batch upload)	Yes	Local	MIT	Version 0.13; August, 2024	None specified
PhyResSE[19]	https://bioinf.fz-borstel.de/mchips/phyresse/	Free	Yes, web portal	No	Remote files and results deleted after 1 month	None specified	Version 1.0; August, 2024	None specified
SAM-TB[20]	https://samtb.uni-medica.com/	Free (users initially limited to 100 analyses but can request more)	Yes, web portal	No	Remote	None specified	Unspecified; August, 2024	None specified
tbpore[21]	https://github.com/mbhall88/tbpore	Free	No	Yes	Local	MIT	Version 0.7.1; August, 2024	None specified
TBProfiler[22]	https://github.com/jodyphelan/TBProfiler	Free	Yes, web portal	Yes	Remote (web portal) or local (CLI)	GPL 3.0	Version 6.4.1; November, 2024	None specified
TBSeqPipe	https://github.com/KevinLYW366/TBSeqPipe	Free	No	Yes	Local	GPL 3.0	Unspecified; August, 2024	None specified
tbtAMR[23]	https://github.com/MDU-PHL/tbtamr	Free	No	Yes	Local	GPL 3.0	Version 1.0.2; November, 2024	ISO15189 standard (Australian National Association of Testing Authorities)

CLI=command-line interface. GUI=graphical user interface. LMICs=low-income and middle-income countries.

Table: **Characteristics of 12 automated *Mycobacterium tuberculosis* whole-genome sequencing analysis pipelines**

For most pipelines, both variant calling errors and catalogue errors contributed to false negative predictions (appendix 1 p 21 and appendix 2 tabs 23–24). The number of variant calling errors per pipelines ranged from two (GPAS) to 98 (MTBseq), and the number of catalogue errors ranged from zero (TBSeqPipe) to 65 (MTBseq). For MTBseq, 63 (64%) of 98 variant calling errors were associated with heteroresistant isolates, which is unsurprising, given that, on default settings, the pipeline does not report minority alleles (although reporting of minority alleles can be enabled with custom settings). The number of false positive predictions ranged from 46 (MTBseq) to 130 (TBProfiler). A large proportion of false positive predictions for MTBseq, Mykrobe, TBProfiler, and TBSeqPipe were due to their classification of variants of uncertain significance (as specified in the WHO catalogue[26]) as resistant (appendix 1 p 22). MAGMA had the highest number of heteroresistant false positive predictions (n=33), resulting from the reporting of resistance for mutations detected at very low allele frequencies (<10%). Processing speed varied substantially between pipelines. TBProfiler was fastest at processing the 1000 sequences (16·9 h), whereas MTBseq took the longest, at 114·9 h (appendix 1 p 26).

See **Online** for appendix 2

gDST accuracies for rifampicin, isoniazid, ethambutol, moxifloxacin, levofloxacin, amikacin, and kanamycin were similar across all 11 Illumina-compatible pipelines for the dataset of 100 Illumina sequences (figure 3; appendix 1 pp 16–17). GenoMycAnalyzer was an outlier, with high false positive rates for several drugs. We identified the reporting of resistance-associated variants at very low allele frequencies (<10%) as the source of these false positive predictions. All 11 pipelines had high false negative rates for bedaquiline, clofazimine, delamanid, and linezolid, with values ranging from 38·5% (95% CI 17·7–64·5) for delamanid (MAGMA) to 100% for bedaquiline and clofazimine (GenoMycAnalyzer and TBSeqPipe; 100·0% [51·0–100·0] for bedaquiline and 100·0% (80·6–100·0) for clofazimine [GenoMycAnalyzer]; 100·0% [43·9–100·0] for bedaquiline and 100·0% [79·6–100·0] for clofazimine [TBSeqPipe]; appendix 1 p 17). False positive rates for the new and repurposed drugs were ≤10% for all pipelines.

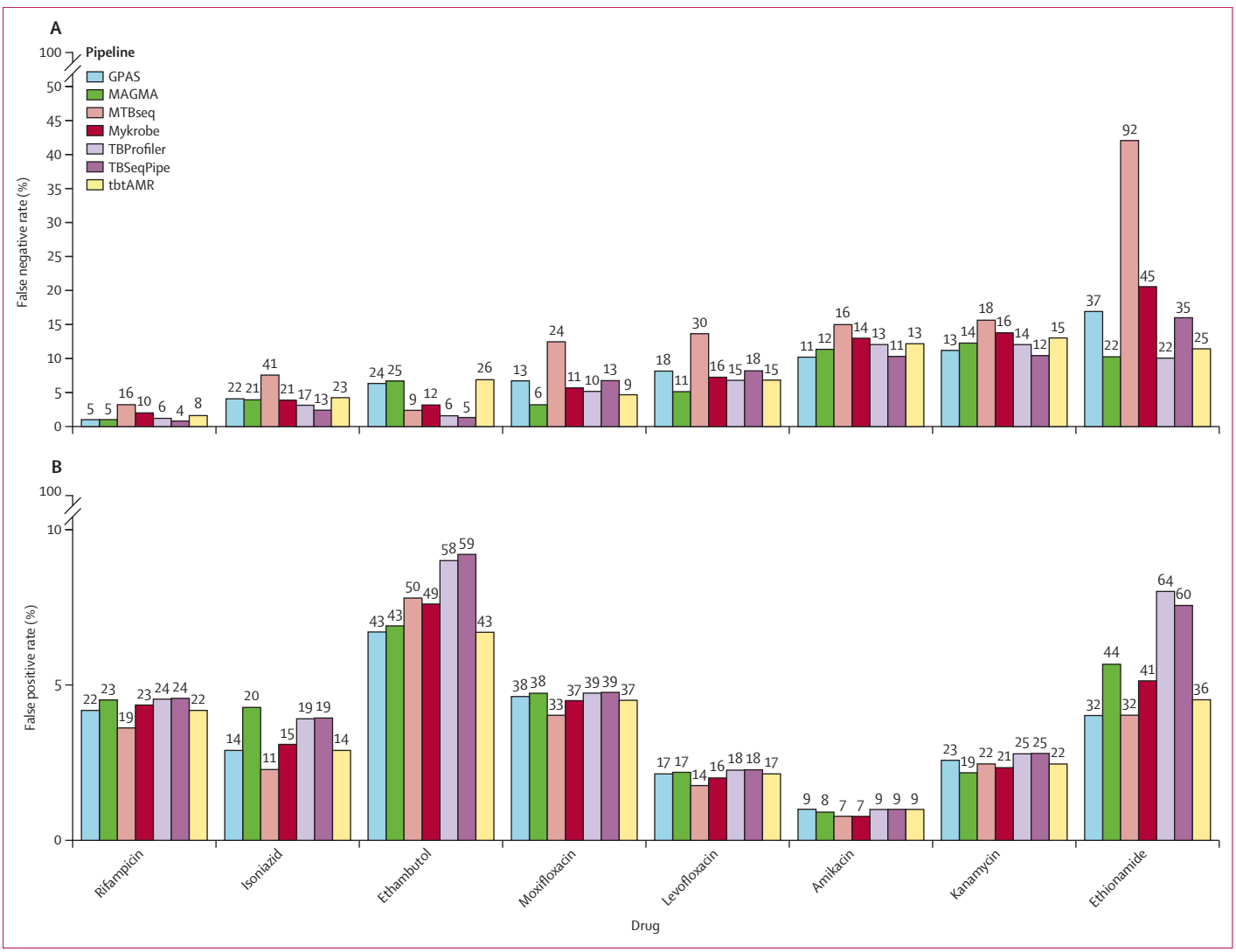

***Figure 2*: False negative rates (A) and false positives rates (B) of genotypic drug susceptibility testing for rifampicin, isoniazid, ethambutol, moxifloxacin, levofloxacin, amikacin, kanamycin, and ethionamide for seven Illumina-compatible *Mycobacterium tuberculosis* whole-genome sequencing analysis pipelines—1000 Illumina sequences dataset**
Values above bars denote absolute numbers of false negative (A) and false positive (B) predictions.

Mykrobe had the lowest overall error rate, at 6·4% (64 errors per 1000 predictions), whereas Gen-TB had the highest error rate at 17·7% (124 per 700 predictions), largely driven by the high proportion of predictions that were failures (84 [12%] per 700 predictions; appendix 1 p 23).

gDST accuracies were similar across Nanopore-compatible pipelines (GPAS, Mykrobe, TBProfiler, and tbpore) and compared well with data on the same samples generated through Illumina platforms (appendix 1 pp 18, 24) but were lower overall than observed with the Illumina-only datasets. As the dataset for the Nanopore-compatible pipelines was limited to just 90 isolates with non-standardised phenotyping, these results should be considered exploratory rather than definitive. Detailed results are provided in appendix 1 (pp 8, 18, 24–25).

None of the Illumina-compatible pipelines incorrectly reported any of the ten non-tuberculous mycobacteria isolates as *M tuberculosis*. For lineage classification, all datasets included isolates from *M tuberculosis* lineages 1–4, with a bias towards lineages 2 and 4 (appendix 1 p 19). Main lineage classification—ie, whether an isolate was assigned as lineage 1, 2, 3, or 4—was concordant across pipelines, with kappa values of 1·00 for the 100 Illumina sequences, 0·99 for the 1000 Illumina sequences, and 1·00 for the 90 Nanopore sequences. Interpipeline disagreement was, however, evident at sublineage resolution, with kappa values of 0·68, 0·66, and 0·84, respectively, for the three datasets. Most pipelines used TBProfiler or Mykrobe for lineage classification, and differences in their typing schemas were the primary source of discrepancy in sublineage

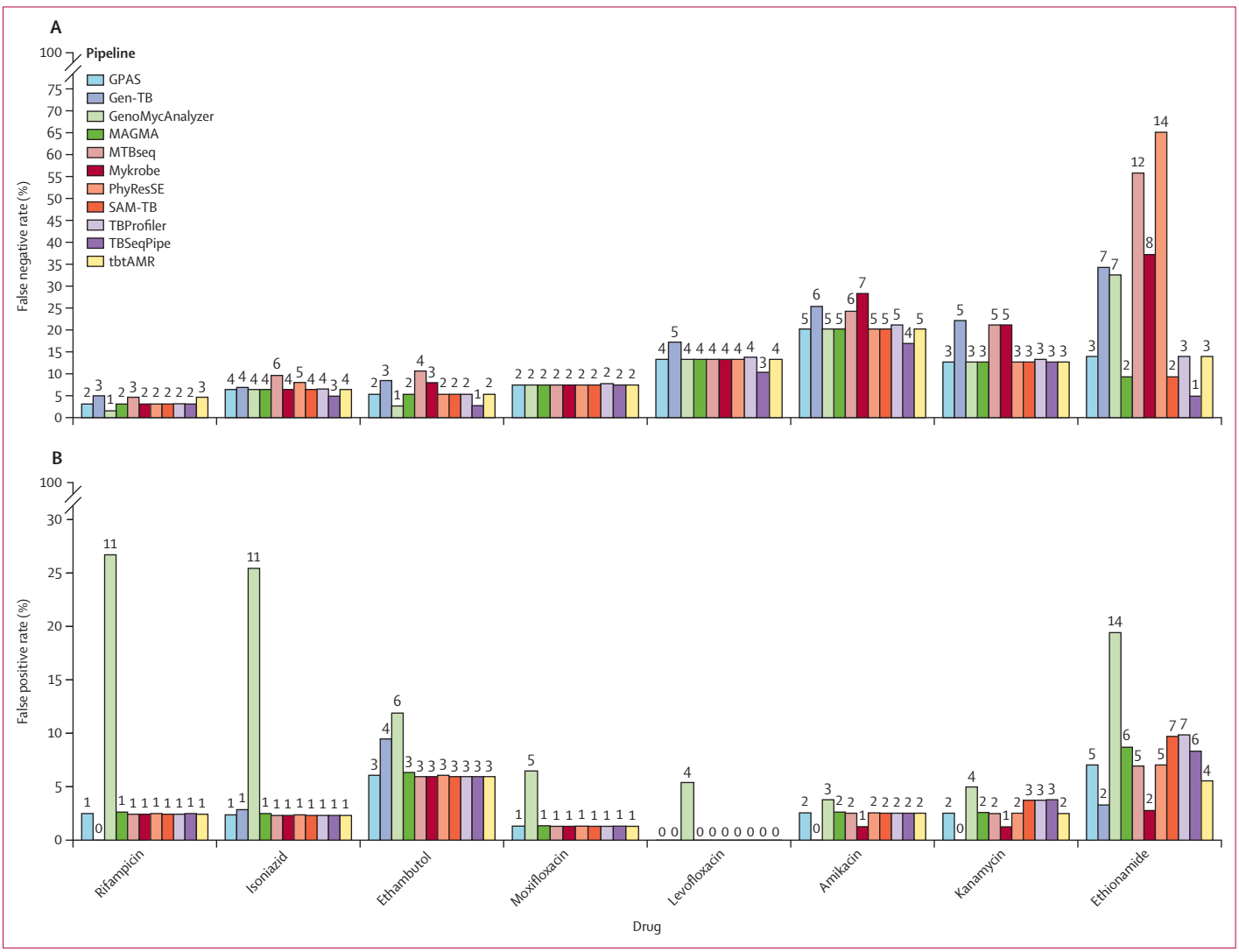

Figure 3: **False negative rates (A) and false positives rates (B) of genotypic drug susceptibility testing for rifampicin, isoniazid, ethambutol, moxifloxacin, levofloxacin, amikacin, kanamycin, and ethionamide with 11 Illumina-compatible *Mycobacterium tuberculosis* whole-genome sequencing analysis pipelines—100 Illumina sequences dataset**
Values above bars denote absolute numbers of false negative (A) and false positive (B) predictions.

classification. These differences appeared to be driven by nomenclature rather than by resolution or SNP calling (appendix 2 tabs 19–21).

We used a fourth dataset consisting of 68 epidemiologically calibrated *M tuberculosis* sequence pairs to evaluate genomic relatedness. Outputs from GPAS, MTBseq, and SAM-TB closely approximated the original findings[12] for this dataset (figure 4), with most distances fewer than five SNPs, and none greater than 12 SNPs, with the exception of two pairs understood to represent infections with distinct strains. TBSeqPipe reported a greater range in SNP differences between linked cases (ranging from one to 65 SNPs) compared with other pipelines, suggesting an over-calling of variants in the genome.

Discussion

We systematically identified 26 automated WGS analysis pipelines for *M tuberculosis* and evaluated 12, some of which were compatible with both Illumina and Nanopore platforms. Our findings will help inform a broad range of users in selecting the most appropriate analysis tool for their needs.

gDST accuracy was similar across most Illumina-compatible pipelines but was substantially lower for the new and repurposed drugs (bedaquiline, clofazimine, delamanid, and linezolid) than it was for the first-line drugs (rifampicin, isoniazid, pyrazinamide, and ethambutol), fluoroquinolones (moxifloxacin and levofloxacin), and injectables (amikacin and kanamycin). Although this finding was unsurprising, given our present understanding of the genomic basis of antituberculosis drug resistance, this lower accuracy remains a concern.[27,28] Particularly concerning is the rapid emergence of resistance to bedaquiline, which threatens recent advances in the development of effective drug regimens.[28,29] A notable strength of gDST compared with standard assays is the capability for rapid updates to resistance mutation catalogues as new

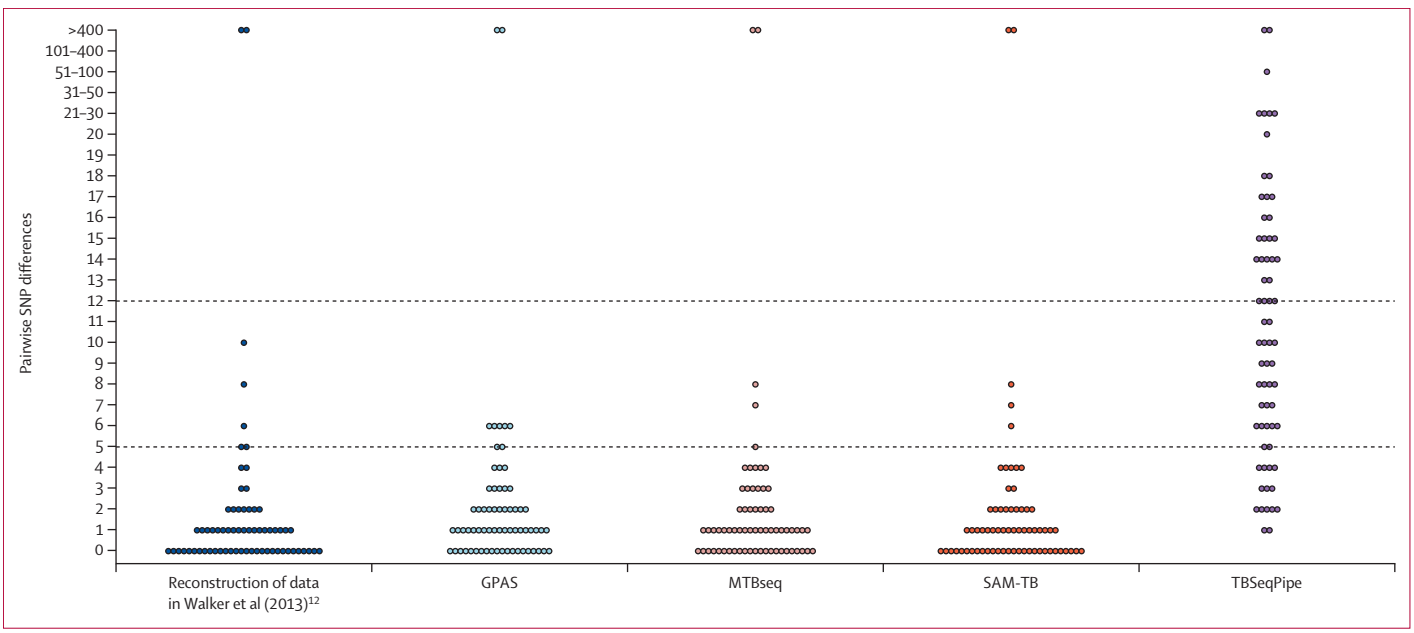

Figure 4: **Differences in numbers of SNPs between sequenced isolates in 68 epidemiologically linked tuberculosis cases reported by four *Mycobacterium tuberculosis* whole-genome sequencing analysis pipelines**

The original study by Walker and colleagues[12] serves as a reference for these results. Dashed horizontal lines represent conventional five-SNP and 12-SNP thresholds. SNP=single nucleotide polymorphism.

genomic insights emerge,[25,26] so results might well improve with time. Accuracy was more variable and generally lower across pipelines for ethionamide than it was for the other drugs, likely due to challenges in interpreting pDST. These challenges stem from the modest MIC increases associated with many ethionamide resistance mutations and the resulting overlap between wild-type and resistant MIC distributions.[30]

Our Illumina sequence test datasets contained few isolates resistant to the new and repurposed drugs and lacked pDST results for pyrazinamide. Moreover, the UKMYC5/6 96-well broth microdilution plates used by the CRyPTIC study[7] to generate the phenotypes are not endorsed by WHO. Nevertheless, the size, standardised sample-processing approach, and global diversity of the CRyPTIC dataset contributed to the strength of this study, and we selected only high confidence phenotypes as identified by CRyPTIC's own rigorous analysis.[7] We expect that this approach will have minimised phenotypic error and contributed to the generalisability of our findings. CRyPTIC did not have clinical outcome data, and data associating specific mutations and patient outcomes remain rare. Further work is required to link genotypic predictions to clinical outcomes, as the ultimate goal of gDST is to guide treatment decisions by accurately predicting drug effectiveness, rather than simply replicating pDST results. Our evaluation did not include direct-from-sample WGS data—an emerging application with clear potential to accelerate diagnosis and transmission inference—due to the absence of publicly available, high-quality datasets containing matched phenotypic results. This sample type poses distinct challenges related to low *M tuberculosis* read depth and high contamination, and only MAGMA is explicitly optimised for this use case.[20,31]

Three of four candidate pipelines clustered sequenced isolates from epidemiologically linked tuberculosis cases within established thresholds, aligning with previous findings for five in-house pipelines.[32] This finding suggests that the relatedness outputs generated from these pipelines might be directly compatible, at least for the SNP differences most relevant to surveillance and outbreak investigation. However, we identified that one pipeline generated notably different results, highlighting the need to interpret results in the context of the specific pipeline used. Similarly, methodological context is paramount when interpreting lineage classification results, especially those at sublineage resolution—a resolution at which commonly used tools yield seemingly discordant results. This lack of consensus might compromise comparability between studies and should ideally be resolved by the tuberculosis research community.

Although gDST accuracy was generally lower in the Nanopore sequence datasets than in the Illumina-only datasets, the consistency observed across pipelines and platforms suggests that differences might be due to phenotypic errors in the 90-isolate dataset, which was not from CRyPTIC, rather than Nanopore's basecalling accuracy. Unfortunately, the absence of publicly available Nanopore-sequenced isolates with corresponding pDST results constrained our ability to comprehensively evaluate gDST accuracy among Nanopore-compatible pipelines. We were also unable to evaluate outputs on genomic relatedness due to the unavailability of epidemiologically calibrated Nanopore sequences. Long-read sequencing for

M tuberculosis still lacks large-scale validation, and efforts to achieve harmonisation between sequencing platforms remain a priority. Expanding the generation and sharing of Nanopore-sequenced *M tuberculosis* samples will be central to these efforts.

The similar accuracy and overlapping range of functionality between most pipelines suggest that the additional non-functional attributes highlighted in this study could guide prospective implementers in their choice of pipeline. Encouragingly, all pipelines were available free of charge, and many offered user-friendly GUIs. However, there was a clear trade-off between user-friendliness and scalability, with CLIs required to process large numbers of sequences—emphasising the importance of training and support, which we did not evaluate in this study. Compared with single-interface pipelines, those providing both GUIs and CLIs might be useful to a wider range of users, as they can accommodate varying levels of user-friendliness and scalability. The data storage strategies used by pipelines also influence their scalability. Not all users in LMICs will have the server capacity to store outputs locally. Remote storage might address this problem but might not be acceptable to national programmes, regardless of the security measures in place. Rapid processing is essential if WGS results for tuberculosis are to inform clinical and public health decisions in a timely manner. Although our analyses were done in a high-performance computing environment that might not reflect conditions in all low-resource settings, the relative processing times of the different pipelines remain informative. In such settings, remote-processing pipelines, which rely on local computational resources for data upload only, could be particularly useful.

Data privacy and security, although not comprehensively reviewed in this study, will become increasingly important considerations as third-party analysis pipelines are scaled up for clinical and public health use. Mycobacterial specimens might be contaminated with human DNA, introducing personally identifiable information and raising regulatory and ethical concerns.[33] Frameworks such as the EU's General Data Protection Regulation and the international ISO 27001 standard specify requirements for data protection and information security. Compliance with these standards will likely be necessary for future regulatory approval.

Ensuring compliance with data protection standards requires ongoing investment, not just in technical measures such as encryption but also in maintaining sound organisational processes and information management systems. Although some pipelines have shown longevity, the numerous unmaintained pipelines excluded from this study and lack of recent updates in some included tools highlight the vulnerability of academic projects reliant on grant funding and the knowledge and expertise of individual contributors. Sustainable funding models are crucial to ensure the long-term viability of these tools.

Our study had several limitations. Resistance prediction and lineage classification outputs might differ according to how variants are identified or which catalogues are used to annotate them. Although we attempted to adjudicate these differences, either factor can be updated, and the performance characteristics we assess here, thus, only represent a cross-section in time. Moreover, we executed pipelines with their default settings, representing the likely behaviour of non-expert users; differences in performance might have been observed with optimised parameters. Only four human-adapted *M tuberculosis* complex lineages were represented in our analyses. Although lineages 1–4 are responsible for most human disease globally, our findings might be less generalisable to settings where other lineages predominate. The core attributes we identified and evaluated beyond accuracy and functionality were restricted to what we could reasonably comment on, rather than being comprehensive. Finally, bioinformatic analysis is only one component of a wider ecosystem required to enable implementation of WGS for *M tuberculosis* in LMICs, which additionally requires sustainable financing, capacity building, and investment in laboratory and computational infrastructure.[34]

We have evaluated numerous automated analysis pipelines that have the potential to increase equity in *M tuberculosis* WGS. As the performance of most pipelines was similar in terms of prediction of resistance, lineage, and relatedness, potential implementers in high-burden, low-resource settings, such as some LMICs, could consider additional features such as cost, user-friendliness, scalability, and privacy and security when deciding on adopting one of these solutions. Progressing these pipelines from research-only tools to fully validated clinical and public health services is a major challenge for the global tuberculosis community and should be considered a priority if the full potential of WGS for tuberculosis control is to be harnessed.

Contributors

RS was responsible for conceptualisation, investigation, data curation, formal analysis, visualisation, and writing the original draft. PWF, RT, and HT were responsible for investigation, data curation, and writing (reviewing and editing). DWC, TEAP, and JAW were responsible for methodology, supervision, and writing (reviewing and editing). TMW was responsible for conceptualisation, investigation, methodology, supervision, and writing (reviewing and editing). RS and TMW verified the underlying data. All authors had full access to all the data in the study and had the final responsibility for the decision to submit for publication.

Declaration of interests

DWC, TEAP, PWF, RT, and HT have previously received financial support from the Ellison Institute of Technology (Oxford, UK), the sponsor of the Global Pathogen Analysis Service. All other authors declare no competing interests.

Data sharing

Accession numbers for all sequences analysed in this study are presented in appendix 2.

Acknowledgments

RS is supported by the Rhodes Trust. This research was funded in whole, or in part, by the Wellcome Trust (214560/Z/18/Z). TMW is a Wellcome Trust

Clinical Career Development Fellow (214560/Z/18/Z). DWC and PWF are supported by funding from the UK National Institute for Health and Care Research (NIHR) Oxford Biomedical Research Centre and the NIHR Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance (NIHR200915), a partnership between the UK Health Security Agency and the NIHR Oxford Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the UK National Health Service, NIHR, the Department of Health and Social Care, or the UK Health Security Agency. During the preparation of this work the authors used ChatGPT-4o to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1 UK Health Security Agency. *Mycobacterium tuberculosis* whole-genome sequencing and cluster investigation handbook. Updated Sept 5, 2022. https://www.gov.uk/government/publications/tb-strain-typing-and-cluster-investigation-handbook/mycobacterium-tuberculosis-whole-genome-sequencing-and-cluster-investigation-handbook (accessed Sept 8, 2024).

2 Andrés M, Van Der Werf MJ, Ködmön C, et al. Molecular and genomic typing for tuberculosis surveillance: a survey study in 26 European countries. *PLoS One* 2019; **14:** e0210080.

3 WHO. Global tuberculosis report, 2024. World Health Organization, 2024.

4 Onywera H, Ondoa P, Nfii F, et al. Boosting pathogen genomics and bioinformatics workforce in Africa. *Lancet Infect Dis* 2024; **24:** e106–12.

5 Aruhomukama D, Galiwango R, Meehan CJ, Asiimwe B. Enhancing genomics and bioinformatics access in Africa: an imperative leap. *Lancet Microbe* 2024; **5:** e410–11.

6 Calero-Cáceres W, Balcázar JL. Leveraging genomic surveillance for public health: insights from Latin America. *Lancet Microbe* 2024; **5:** 100919.

7 The CRyPTIC Consortium. A data compendium associating the genomes of 12,289 *Mycobacterium tuberculosis* isolates with quantitative resistance phenotypes to 13 antibiotics. *PLoS Biol* 2022; **20:** e3001721.

8 Dorai-Raj S. Binom: binomial confidence intervals for several parameterizations. R package version 1.1–1.1. https://cran.r-project.org/web/packages/binom/index.html (accessed Sept 8, 2024).

9 Doebler P, Sousa-Pinto B. mada: meta-analysis of diagnostic accuracy. R package version 0.5.11. 2022. https://cran.r-project.org/web/packages/mada/index.html (accessed Sept 8, 2024).

10 Hall MB, Lima L, Coin LJM, Iqbal Z. Drug resistance prediction for *Mycobacterium tuberculosis* with reference graphs. *Microb Genom* 2023; **9:** mgen001081.

11 Gamer M, Lemon J, Fellows I, Singh P. irr: various coefficients of interrater reliability and agreement. R package version 0.84.1. 2019. https://cran.rproject.org/web/packages/irr/index.html (accessed Sept 8, 2024).

12 Walker TM, Ip CLC, Harrell RH, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect Dis* 2013; **13:** 137–46.

13 Westhead J, Baker CS, Brouard M, et al. Enhancement and validation of the antibiotic resistance prediction performance of a cloud-based genetics processing platform for Mycobacteria. *bioRxiv* 2024; published online Nov 8. https://www.biorxiv.org/content/10.1101/2024.11.08.622466v1 (preprint).

14 Gröschel MI, Owens M, Freschi L, et al. GenTB: a user-friendly genome-based predictor for tuberculosis resistance powered by machine learning. *Genome Med* 2021; **13:** 138.

15 Kim D, Shin JI, Yoo IY, et al. GenoMycAnalyzer: a web-based tool for species and drug resistance prediction for Mycobacterium genomes. *BMC Genomics* 2024; **25:** 387.

16 Heupink TH, Verboven L, Sharma A, et al. The MAGMA pipeline for comprehensive genomic analyses of clinical *Mycobacterium tuberculosis* samples. *PLoS Comput Biol* 2023; **19:** e1011648.

17 Kohl TA, Utpatel C, Schleusener V, et al. MTBseq: a comprehensive pipeline for whole genome sequence analysis of *Mycobacterium tuberculosis* complex isolates. *PeerJ* 2018; **6:** e5895.

18 Hunt M, Bradley P, Lapierre SG, et al. Antibiotic resistance prediction for *Mycobacterium tuberculosis* from genome sequence data with Mykrobe. *Wellcome Open Res* 2019; **4:** 191.

19 Feuerriegel S, Schleusener V, Beckert P, et al. PhyResSE: a web tool delineating *Mycobacterium tuberculosis* antibiotic resistance and lineage from whole-genome sequencing data. *J Clin Microbiol* 2015; **53:** 1908–14.

20 Yang T, Gan M, Liu Q, et al. SAM-TB: a whole genome sequencing data analysis website for detection of *Mycobacterium tuberculosis* drug resistance and transmission. *Brief Bioinform* 2022; **23:** bbac030.

21 Hall MB, Rabodoarivelo MS, Koch A, et al. Evaluation of nanopore sequencing for *Mycobacterium tuberculosis* drug susceptibility testing and outbreak investigation: a genomic analysis. *Lancet Microbe* 2023; **4:** e84–92.

22 Phelan JE, O'Sullivan DM, Machado D, et al. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med* 2019; **11:** 41.

23 Horan KA, Viberg L, Ballard SA, et al. Bringing TB genomics to the clinic: a comprehensive pipeline to predict antimicrobial susceptibility from genomic data, validated and accredited to ISO standards. *bioRxiv* 2023; published online Nov 6. https://doi.org/10.1101/2023.11.04.565651.

24 Coll F, McNerney R, Guerra-Assunção JA, et al. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat Commun* 2014; **5:** 4812.

25 Walker TM, Miotto P, Köser CU, et al. The 2021 WHO catalogue of *Mycobacterium tuberculosis* complex mutations associated with drug resistance: a genotypic analysis. *Lancet Microbe* 2022; **3:** e265–73.

26 WHO. Catalogue of mutations in *Mycobacterium tuberculosis* complex and their association with drug resistance, 2nd edn. World Health Organization, 2023.

27 Nimmo C, Millard J, Faulkner V, Monteserin J, Pugh H, Johnson EO. Evolution of *Mycobacterium tuberculosis* drug resistance in the genomic era. *Front Cell Infect Microbiol* 2022; **12:** 954074.

28 Van Rie AV, Walker T, De Jong B, et al. Balancing access to BPaLM regimens and risk of resistance. *Lancet Infect Dis* 2022; **22:** 1411–12.

29 Barilar I, Fernando T, Utpatel C, et al. Emergence of bedaquiline-resistant tuberculosis and of multidrug-resistant and extensively drug-resistant *Mycobacterium tuberculosis* strains with rpoB Ile491Phe mutation not detected by Xpert MTB/RIF in Mozambique: a retrospective observational study. *Lancet Infect Dis* 2024; **24:** 297–307.

30 WHO. WHO operational handbook on tuberculosis. Module 3: diagnosis. Rapid diagnostics for tuberculosis detection Annex C. Technical manual for culture-based drug susceptibility testing of anti-tuberculosis drugs used in the treatment of tuberculosis. World Health Organization, 2024.

31 Goig GA, Cancino-Muñoz I, Torres-Puente M, et al. Whole-genome sequencing of *Mycobacterium tuberculosis* directly from clinical samples for high-resolution genomic epidemiology and drug resistance surveillance: an observational study. *Lancet Microbe* 2020; **1:** e175–83.

32 Jajou R, Kohl TA, Walker T, et al. Towards standardisation: comparison of five whole genome sequencing (WGS) analysis pipelines for detection of epidemiologically linked tuberculosis cases. *Euro Surveill* 2019; **24:** 1900130.

33 Constantinides B, Hunt M, Crook DW. Hostile: accurate decontamination of microbial host sequences. *Bioinformatics* 2023; **39:** btad728.

34 Meehan CJ, Goig GA, Kohl TA, et al. Whole genome sequencing of *Mycobacterium tuberculosis*: current standards and open issues. *Nat Rev Microbiol* 2019; **17:** 533–45.