Tiled Amplicon Sequencing Enables Culture-free Whole-Genome Sequencing of Pathogenic Bacteria From Clinical Specimens

Chaney C Kalinich[1 *], Freddy L Gonzalez[2 *], Alice Osmaston[4,5], Mallery I Breban[1], Isabel Distefano[1], Candy Leon, Patricia Sheen, Mirko Zimic, Jorge Coronel, Grace Tan, Walter Solano[5], Jimena Belén Ráez[5], Orchid M Allicock[1], Chrispin Chaguza[1], Anne L Wyllie[1], Matthew Brandt[6], Daniel M. Weinberger[1,3], Benjamin Sobkowiak[4,3], Ted Cohen[1,3], Louis Grandjean[4,5], Nathan D Grubaugh[1,2,3], Seth N Redmond[1,7]

1 Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, Connecticut, USA

2 Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut, USA

3 Public Health Modeling Unit, Yale School of Public Health, New Haven, Connecticut, USA

4 University College, London

5 Cayetano Heredia University, Peru

6 Yale School of Medicine, New Haven, Connecticut, USA

7 Yale Institute for Global Health, Yale University, New Haven, Connecticut, USA

* these authors contributed equally to this work

Abstract

Pathogen sequencing is an important tool for disease surveillance and demonstrated its high value during the COVID-19 pandemic. Viral sequencing during the pandemic allowed us to track disease spread, quickly identify new variants, and guide the development of vaccines. Tiled amplicon sequencing, in which a panel of primers is used for multiplex amplification of fragments across an entire genome, was the cornerstone of SARS-CoV-2 sequencing. The speed, reliability, and cost-effectiveness of this method led to its implementation in academic and public health laboratories across the world and adaptation to a broad range of viral pathogens.However, similar methods are not available for larger bacterial genomes, for which whole-genome sequencing typically requires *in vitro* culture. This increases costs, error rates and turnaround times. The need to culture poses particular problems for medically important bacteria such

as *Mycobacterium tuberculosis,* which are slow to grow and challenging to culture. As a proof of concept, we developed two novel amplicon panels for *Streptococcus pneumoniae* and *Mycobacterium tuberculosis*, which enabled recovery of whole bacterial genomes without culturing. Applying our amplicon panels to clinical samples, we show the ability to classify pathogen subgroups and to reliably identify markers of drug resistance. Development of this work in clinical settings has the potential to tailor disease interventions and treatment regimes for these high priority pathogens.

Introduction

In recent years, whole-genome amplicon sequencing has been adopted as a standard technique for genomic surveillance of infectious disease. Initially developed for genomic surveillance of the 2016 Zika epidemic [1], where low viraemia had precluded direct sequencing of clinical samples even where infection had been confirmed. Amplicon sequencing uses multiplex PCR of tiled overlapping regions of a target genome to recover whole genomes from samples of low concentration or complex backgrounds. This has proven particularly useful for sequencing remnant samples from diagnostic tests, and its use in the 'artic' protocol for sequencing SARS-CoV-2 [2] has led to it being deployed in thousands of public health laboratories around the world, facilitating true global surveillance of viral dynamics [3]. The ease, reliability and low cost of this approach has seen its adaptation to a broad range of viral pathogens both in respiratory disease [4,5] and beyond [6,7]. However, viral infections are by no means the only cause of global morbidity; more than half of all infectious disease-related deaths are caused by 33 bacterial pathogens, with *Streptococcus pneumoniae* alone estimated to be responsible for more than 800,000 deaths per year [8].

As with viral pathogens, sequencing of bacteria can enable reconstruction of transmission chains for targeting interventions and routine surveillance of pathogen diversity for vaccine design or monoclonal antibody targeting. It can also be uniquely informative for the detection of drug resistance in bacteria, allowing insights into the horizontal transfer of antimicrobial resistance genes and potentially enabling more effective tailored treatment regimes [9]. However, the relatively laborious process of isolating and culturing patient samples means this is typically only performed where a small number of samples can be highly informative, such as in outbreaks of food-borne pathogens [10–12], or drug-resistant nosocomial infections [13–15]. This kind of approach is far less appropriate for a bacterium with high rates of commensal and asymptomatic transmission such as *S. pneumoniae*, and it would be entirely cost-prohibitive in the 24 low- and middle-income countries (LMICs) in which it is the leading cause of death [8]. A disease such as tuberculosis (TB) has similarly high burden and low detection rates. In addition, *Mycobacterium tuberculosis* has low genomic diversity, making traditional approaches, such as restriction fragment length polymorphism (RFLP), spoligotyping, and variable number tandem repeat (VNTR) less precise [16]. A notoriously slow growth rate means it can take weeks to detect, let alone sequence, a TB infection [16]. Though whole genome sequencing (WGS) of *M. tuberculosis* has demonstrated clear application to detecting superspreading [17], or distinguishing recrudescence from reinfection [18], the difficulties of culturing *M. tuberculosis* means these studies have typically been performed retrospectively. The use of WGS to inform outbreak investigations as they occur has been limited to high-resource settings such as the United Kingdom [19]. Adapting the techniques used for viral pathogens to bacterial pathogens, enabling rapid culture-free sequencing from minimal input volumes and the use of remnant tests and other passive surveillance techniques, could be transformative for bacterial genomic epidemiology.

We present here the first use of amplicon-based WGS for the sequencing of two bacterial pathogens of major public health importance. We have designed tiling amplicon schemes for *S. pneumoniae* serotype 3 and *M. tuberculosis*. These assays are able to generate complete genome coverage from samples with minimal input concentrations without any requirement for bacterial culturing. We show recovery of genomic data from a broad range of sample types, including saliva, sputum, nasopharyngeal swabs and remnant diagnostic tests, and further show that this genomic data can reliably perform *in-silico* lineage or serotype assignment to enable the surveillance of bacterial transmission dynamics. We show that our TB amplicon panel can be applied directly to sputum samples to identify clinically relevant phenotypes such as antimicrobial susceptibility within days of sample collection, and can detect resistance loci that were not found by rapid diagnostics. We hope that this work will not only generate opportunities for future genomic epidemiology of *S. pneumoniae* and *M. tuberculosis*, but will also provide a roadmap for the development of amplicon sequencing for other clinically important bacterial pathogens.

Results

In silico **predictions indicate broad applicability of amplicon schemes across clades**

In order to design primer schemes with efficient amplification of diverse target sequences, we downloaded a selection of whole-genome sequences available on public repositories for both *M. tuberculosis* (n=489, **Supplemental file 1a**) and *S. pneumoniae* (n=490, **Supplemental file 1b**). For *S. pneumoniae*, we assembled these sequences into a 'metaconsensus' sequence, a reference-guided core genome with SNPs and indels replaced with 'N'. PrimalScheme was run on the output of this, in order to design primers which cover the core *S. pneumoniae* genome and avoid variant sites. Because *M. tuberculosis* has very little within-species diversity outside of the repetitive hypervariable PE/PPE/PGRS regions, PrimalScheme was run directly on the H37Rv reference genome after masking PE/PPE/PGRS regions and sites with known resistance-related polymorphisms to avoid primers being designed at these loci.

For both pathogens, we selected a small number of genetically diverse sequences from the larger set of publicly-available sequences to predict coverage beyond the sequences used for primer panel design. As expected, predicted amplicon coverage was highest against the strains used for panel design (*Sp*:CC180: 98.93%; *Mt*/H37Rv: 94.31%) - *M. tuberculosis* coverage is reduced due to omission of PE/PPE regions from the design, which account for 8-10% of the genome [20]. However, predicted coverage in *M. tuberculosis* remained high across all 7 lineages (>= 94.23%) and in the *M. canetti* outgroup (89.44%), while *S. pneumoniae* fell sharply across the clade (>=89.44) and in the *S. mitis* outgroup (32.18%)

(Figure 1). Average nucleotide identity fell less across the clade for *M. tuberculosis* (99.98-99.27%) than *S. pneumoniae* (98.76-92.51%), however pangenome size and similarity was markedly different between the two species, with *M. tuberculosis* having a smaller relative pangenome (4,335 total genes and a mean genome size of 4,067 genes, a ratio of approximately 1.07, compared to 3,942 total genes and a mean genome size of 2,071 genes for *S. pneumoniae,* a ratio of approximately 1.90) and, as a result, far more sharing of genes with the reference strain (4,002-4,050) than *S. pneumoniae* (1,705-1,793). Our findings suggest panel applicability is largely affected by genome rearrangement rather than increases in genetic distance.

Amplicon sequencing enables recovery of whole genomes from diverse and minimal-input bacteria samples

To determine our ability to enrich target genomes from within diverse sample sources, for *S. pneumoniae*, we sequenced DNA from cultured isolates, nasopharyngeal (NP) swabs, saliva samples, and culture-enriched saliva and NP swabs using our amplicon panel and standard metagenomic sequencing. For *M. tuberculosis*, we sequenced DNA from cultured isolates and sputum samples using the same amplicon sequencing workflow, with and without adding amplicon panel primers, as well as standard metagenomic sequencing.

For *S. pneumoniae*-positive samples, despite high species diversity in each sample type, increases in the proportion of reads mapping to the target genome were seen for both saliva and nasopharyngeal swabs compared to standard metagenomic sequencing, alongside an additional increase in related *Streptococcus* species **(Figure 2a-b)**; this is particularly noticeable within saliva samples, which are expected to have high complements of *S. oralis* and *S. mitis*. *S. pneumoniae* read recruitment was high among cultured isolates regardless of amplification protocol.

Comparisons of amplified and unamplified *M. tuberculosis*-positive sputum samples demonstrated dramatic increases in coverage for amplified samples as compared to the same samples without amplification **(Figure 2c-d)**. While only 2/10 of unamplified samples achieved more than 75% coverage, 9/10 of the amplified samples sequenced above this threshold, with 7 of those generating more than 95% coverage. The remaining sample achieved 33% coverage amplified, and negligible coverage unamplified. Metagenomic sequencing indicated lower overall species diversity for TB sputum samples, yet successful amplification from samples containing both commensal and pathogenic bacteria including *Streptococcus, Pseudomonas, Actinomycetes* and *Schaalia spp*.

We assessed the limits of detection for each amplicon panel by sequencing serial 10-fold dilutions of 6 cultured samples of each bacteria using both amplified and unamplified sequencing approaches. For *M. tuberculosis*, high genome coverage (>95%) was observed in all amplified samples above 100 genome copies per microlitre (GC/µL), compared to 10,000 GC/µL for unamplified samples **(Fig S1)**.

Amplicon derived data enables phylogenetic classification and population delineation of *M. tuberculosis*

Lineages of *M. tuberculosis* were called with the Mykrobe package [21], which assigned all samples to lineages 2 (sublineage 2.2.1) or 4. Mykrobe performed equally well in high coverage samples, regardless of whether these were derived from cell culture or sputum. We derived maximum likelihood phylogenies using IQ-tree [22] including all sequenced specimens and the broad reference set of *M. tuberculosis* sequences used for primer design (**Supplemental Figure 2; Supplemental file 1a**). In all cases the primary lineage predicted by Mykrobe aligned with lineages from a maximum-likelihood tree, though in some cases secondary lineages were predicted based on minor variants which did not concord with the ML tree.

Lineage calling for *S. pneumoniae* was largely unsuccessful. Culture-derived samples could be assigned to serotype using either PneumoKITy [23] or PopPunk [24], however none of the lineage callers (PneumoKITy [23], PopPunk [24], SRST2 [25]) were able to assign lineages to any of the direct clinical samples from saliva or nasopharyngeal swabs. While our data did indicate coverage of the 7 major *S. pneumoniae* housekeeping genes (*aroE, gdh, gki, recP, spi, xpt, ddl*) lineage predictions may have been impaired by the high concentration of commensal bacteria in the enriched samples.

Direct sputum sequencing for TB detects markers of antimicrobial resistance to first-line therapies

For *M. tuberculosis*, sequencing data was high-quality enough to produce a prediction for all template dilutions from cultured isolates with at least 10 GE/µL starting quantity (**Figure 3, Supplemental Figure 3**). While we do not have access to phenotypic susceptibility results for these isolates, predictions were internally consistent for all template dilutions above 100 GE/µL (though there was some variability between partial vs full resistance calling) with the exception of streptomycin. DNA was extracted directly, without culturing, from 60 sputum specimens with a range of acid-fast bacilli semi-quantitative measurements (e.g., 1+ to 3+); sequencing data was high-quality enough to produce a drug susceptibility prediction for 53/60 sputum specimens. Of the 7 specimens which failed, (Yale-TB121, Yale-TB123, Yale-TB149, Yale-TB150) had starting quantities (following extraction) below 10 GE/uL. None of the other 3 (Yale-TB126, Yale-TB139, Yale-TB148) have GenXpert results available as comparison. Several different extraction methods were used (detailed in **Supplemental table S1b**) as it was not clear what method would perform best; all 20 specimens extracted with the final protocol, which included a NALC-NaOH treatment to deplete non-mycobacterial DNA, had adequate data to predict resistance.

In-silico antibiotic resistance screening in *S. pneumoniae* identified resistance to several second-line and broad-spectrum antibiotics, including Lincosamides, Macrolides, and Fluoroquinolones in 9/9 culture isolate samples, 7/9 culture enriched samples, 14/15 saliva samples, and 0/3 nasopharyngeal samples. Percent coverage and percent identity toward resistance genes (*RlmAII*, *patA*, *patB*, and *pmrA*) ranged from 76.33% to 100% (mean = 94.85%) and 99.24% to 100% (mean 93.87%) for cultured isolates, 75.04% to 100% (mean 90.6%) and 78.43% to 88.93% (mean 83.72%) for culture-enriched samples, and 75.21% to 100% (mean 93.91%) and 82.84% to 99.38% (mean 92.57%) for saliva samples, respectively. On average, samples contained at least 3 resistance genes (**Supplemental Table 3)**. We identified several virulence factors including capsular polysaccharides, many of which are associated with TIGR4 (Serotype 4) and *Streptococcus pyogenes*, suggesting prior capsular switching and horizontal gene transfer events, highlighting the ability of amplicon sequencing to pick up on the genetic diversity and evolutionary adaptability of *S. pneumoniae*.

Discussion

Tiled amplicon sequencing of pathogens has proven extremely useful for reconstructing disease spread and gaining insight into transmission patterns for a variety of viruses [26]. The 2020 SARS-CoV-2 pandemic stimulated a global effort to adopt these methods and use genomics to track and monitor the virus; however, it has not previously been applied to the significantly larger and often more complex genomes of bacteria. Our work here, in which we have successfully used a tiled amplicon approach to sequence two pathogenic bacteria from specimens with minimal input DNA and demonstrated the ability to identify clades and markers of drug resistance, could have a major impact on disease control for these two species.

Both *S. pneumoniae* and *M. tuberculosis* are pathogens of prime public health importance. *S. pneumoniae* is responsible for more than 800,000 deaths per year, with the majority of these resulting from respiratory tract infections in children under five [8], and vaccine design is guided by ongoing genomic sequencing [27]. Prior to the Covid-19 pandemic, TB was the world's leading cause of death from a single infectious agent, causing more than a million deaths per year [28]. Despite the availability of vaccines, treatment, and significant funding [29], we continue to miss WHO targets for reductions in TB incidence and death by wide margins, indeed cases have risen worldwide over the past 2 years [30].

Antimicrobial resistance is a critical issue in treating and controlling TB, due to the prevalence of resistance to first line drugs and the length, cost, and complexity of treatment regimes [31]. Despite the introduction of shorter regimens, the time taken to find an effective treatment can be long, and incomplete treatment remains a problem [32]. For this reason, the WHO now recommends the use of targeted sequence-based diagnostics for rapid drug susceptibility testing for patients who are at high risk of, or have already experienced, treatment failure [33]. However designing such an assay is not simple; more than 40 separate loci, each containing numerous individual mutations, have been implicated in drug resistance [34], and uncertainty can be higher for new or second line drugs [35]. Whole-genome sequencing works around these limitations of targeted amplicon sequencing. As data are being generated across the entire genome, drug susceptibility predictions can be improved and expanded bioinformatically as new genetic markers are discovered without updating primers, unlike existing targeted sequencing-based diagnostics.

The required time, infrastructure, and costs for tiled amplicon sequencing are almost identical to targeted amplicons; the additional data generated through WGS can be used along with phenotypic drug susceptibility to expand our understanding of the genetic markers of drug resistance, especially for third-line or novel drugs, increasing the accuracy of predictions over time [20].

Whole genome sequencing obviates the need to design a targeted assay and can also return resistance predictions within days of a positive culture. However, the requirement of most existing WGS approaches to first grow a culture sample means that the overall sample-to-sequence turnaround time for *M. tuberculosis* is measured in weeks or months [19] and significant biases can be introduced during the culturing process itself [20]. Direct WGS without culture does not consistently produce data of high enough quality for resistance prediction or thorough epidemiologic investigation [36,37], is limited to specimens with a high bacterial load [37], or relies on expensive techniques such as hybrid capture [38]. We have demonstrated tiled amplicon sequencing directly from sputum specimens, without culture, can be used to make accurate drug susceptibility predictions and lineage assignments for the majority (53/60) of specimens, unlike prior whole-genome sequencing approaches [19,36–39]. For a notoriously slow-growing organism such as *M. tuberculosis*, eliminating this step reduces the time from sample collection to genome from weeks to days. Not only could patients receive appropriate antibiotics sooner, but genomic epidemiology could be used in real-time to inform outbreak investigations [40,41] and public health measures to reduce spread [42,43].

Despite more consistent coverage across *in vitro* and *in silico* predictions, gaps remain in our coverage of the *M. tuberculosis* genome in the PE/PPE regions. While these are frequently omitted from *M. tuberculosis* analyses, increasing evidence of functions in host cell invasion [44] and importance for vaccine design [45] suggest inclusion of these regions in future iterations of this amplicon panel would be a significant improvement.

Nevertheless the comparison between our results for *S. pneumoniae* and *M. tuberculosis* is instructive. Neither panel showed high rates of amplicon dropout when faced with targets which had drifted from the reference strain (a regular issue with viral amplicon panels). However, *in silico* predictions suggest a weaker applicability of the amplicon panel in species which undergo significant levels of recombination and horizontal gene transfer, and our inability to reliably recover serotype and resistance loci in *S. pneumoniae* supports this conclusion. Indeed, *S. pneumoniae* exhibits extremely high levels of horizontal gene transfer, not only within the species, but also with frequently co-occurring commensal bacteria such as *S. mitis* and *S. oralis* [46,47]. This species may simply not be a suitable target for this approach, where metagenomic or hybrid capture-based sequencing may be more appropriate.

Faced with both drift and genomic rearrangement, designing primers that target conserved motifs will rely upon databases of previously sequenced genomes to allow us to determine circulating genetic diversity. Rapid improvements in sequencing and assembly technology have generated vast databases of assembled genomes; while these resources are not comprehensive, their bias towards improved representation of species of clinical interest [48] suggests this will not be a limiting factor in panel design.
An alternative consideration is targeting bacteria which do not undergo significant levels of horizontal gene transfer, and the ratio of genome to pangenome size is likely to be a key metric for our ability to design an amplicon panel. This ratio is highly sensitive to the diversity of habitats in which the pathogen

is found: free living or commensal species gain particularly large pangenomes to enable adaptation to diverse environments; intracellular pathogens show strong purifying selection, low effective population sizes and low genome:pangenome ratios [49]. *M. tuberculosis*, an obligate pathogen and intracellular bacterium which has been extensively sequenced [50], has little horizontal gene transfer, and remains a major threat to human life [30], may be archetypal, however other intracellular pathogens such as *Yersinia pestis*, *Listeria monocytogenes*, *Legionella pneumophila*, and *Chlamydia trachomatis* are suitable targets.

The widespread use of tiled amplicon sequencing for pathogen genomics during the Covid-19 pandemic has ensured that this method is trusted, understood, and easily implemented in academic and public health laboratories worldwide. As the focus now turns to adapting this capacity to other public health threats [3], it is important to prioritize the development of tools for global priority pathogens that can be implemented in the regions suffering the greatest burden. Genomic surveillance of TB has demonstrated capacity to guide TB interventions in high income countries [17,18]; the reductions in cost and turnaround time afforded by tiled amplicon sequencing could enable this to be implemented in LMICs with high TB burden. Just four countries (India, Bangladesh, Indonesia, Democratic Republic of the Congo) account for over half of all TB deaths; all have seen prior in-country amplicon sequencing of SARS-CoV-2 [51–54] suggesting a ready capacity for tiling amplicon sequencing of *M. tuberculosis*. Extensive use of alternative sequencing methods such as Oxford Nanopore in these regions [51,53,55] suggest adaptation to cheaper and more portable sequencing platforms may further increase surveillance capacity.
Barriers to clinical application are necessarily higher [56]; if diagnostics and resistance prediction are to be used to tailor treatment regimes it is vital that they can be shown to work reliably in a range of likely scenarios: paucibacillary infections; mixed *M. tuberculosis* strains; mixed *M. tuberculosis* and non-tuberculosis mycobacteria; partial and incomplete resistance. Despite this complex landscape, the capacity of culture-free *M. tuberculosis* sequencing to allow early diagnosis and resistance detection could be transformative. Not only by increasing completion rates of TB treatment at the patient level [57], but by preventing the further transmission of MDR-TB at the population level [58,59]. A comprehensive

evaluation of culture-free sequencing methods in a clinical environment should be a priority for TB control.

Materials and Methods

Ethics statement

All specimens were discarded and de-identified specimens used previously for diagnostic testing or IRB-approved human subjects research in accordance with Yale University IRB-exempt protocol #2000033281. *S. pneumoniae* specimens were remnant specimens collected from study participants enrolled and sampled in accordance with the Yale University Humans Investigation Committee-approved protocol #2000027690. *M. tuberculosis* specimens from Moldova were remnant specimens collected from study participants enrolled and sampled in accordance protocol #2000023071 approved by Yale University Human Investigations Committee and the Ethics Committee of Research of the Phthisiopneumology Institute in Moldova. *M. tuberculosis* specimens from Peru were remnant specimens collected from study participants enrolled in accordance with protocol #204749 approved by the Institutional Committee on Research Ethics at Cayetano Heredia University, Peru.

Primer design

We downloaded all available *S. pneumoniae* Serotype 3 contigs (n=490; **Supplemental file 1b**) from the Global Pneumococcal Sequencing (GPS) database [60] on 02FEB2023. We downloaded raw reads for *M. tuberculosis* sequences from a previously described globally representative dataset [50] (n=489; **Supplemental file 1a**) from the European Nucleotide Archive (ENA) at EMBL-EBI. For both targets, we downloaded complete reference genomes from the National Center for Biotechnology Information (NCBI) GenBank (OXC141; accession NC_017592 and H37Rv; accession NC_000962.3).

For *M. tuberculosis*, variants were called against the reference using Snippy and time-resolved maximum-likelihood tree was built using our variant call file, along with sample data generated from Augur (v.22.4.0), IQ-Tree (v.2.23), and TreeTime (v.0.10.1). Representative sequences (n=6) were selected from across this tree using Parnas (v.0.1.4), to cover >50% of the expected overall diversity. We used these representatives to create an *M. tuberculosis* core genome assembly using Snippy.

For *S. pneumoniae*, consensus genome sequences were generated (n=4) with Snippy (v.4.6.0).

Tiled primer schemes (target amplicon size 2kb) were designed for both *S. pneumoniae* and *M. tuberculosis* (excluding PE/PPE and repeat regions) using PrimalScheme [1]. Primers were ordered at 100uM and 200uM in IDTE for *S. pneumoniae* and *M. tuberculosis*, respectively. Primer pools consisted of an equal volume of each primer and were used for amplification without further dilution.

Clinical specimens

S. pneumoniae samples consisted of DNA extracted from raw saliva (15), nasopharyngeal swabs in viral transport media (VTM) (6), culture-enriched bacteria (16), and cultured pure isolates (9). All saliva specimens had a paired cultured specimen cultured from the saliva (either culture enriched or cultured isolate); six also had a paired nasopharyngeal swab collected simultaneously from the same patient, three of which were sequenced with and without amplification. A full list of *S. pneumoniae* samples and descriptions can be found in **Supplemental Table S1a**. DNA was extracted from 200uL of each sample using the MagMAX Ultra viral/pathogen nucleic acid isolation kit (Thermo Fisher Scientific) using a KingFisher Apex instrument (Thermo Fisher Scientific) and quantified using two qPCR primer/probe pairs, *lytA* [61] and *piaB* [62] as described previously [63].

M. tuberculosis samples consisted of DNA extracted from positive solid or liquid cultures from sputum and DNA extracted directly from sputum specimens. Extracts from culture consisted of remnant specimens from a prior study in Moldova, where sputum specimens were tested at a number of diagnostic centers in Moldova by microscopy, Xpert, and culture and positive cultures sent to the National TB Reference Laboratory in Chisnau for extraction by the cetyltrimethylammonium bromide (CTAB) method as described previously [42]. Extracts from sputum consisted of specimens collected in Peru after routine diagnostics had been carried out and TB confirmed. In order to test the efficiency of different methods for extracting DNA from sputum, each specimen was split into two and processed with two different protocols. A total of 30 unique sputum specimens were processed with two protocols each, and a total of

6 different protocols were tested. A full list of all *M. tuberculosis* samples and the extraction methods used can be found in **Supplemental Table S1b**, and a detailed description of extraction methods can be found in **Supplemental Methods 1**. Following extraction, DNA was quantified with a mycobacterium tuberculosis-complex specific, fluorescence-based real-time PCR assay on the Bio-Rad CFX96 instrument [64].

Metagenomic sequencing

For *S. pneumoniae*, 1-3ng of each sample (up to 4uL for samples which were undetectable) and a negative template control (4uL H2O) underwent tagmentation for 5 minutes followed by a magnetic bead cleanup. Then, samples were amplified with Nextera dual-index adapters followed by a second magnetic bead cleanup. Each sample was quantified with a Qubit fluorometer and 5ng of each library were pooled together (up to 4uL for undetectable samples). The pooled libraries underwent a final 0.7X bead clean up, then were quantified on a Qubit fluorimeter and quality and fragment distribution verified using an Agilent Bioanalyzer. For *M. tuberculosis*, samples were prepared as described for amplicon sequencing, but with the addition of sterile water in place of PCR primer pools for both amplification reactions.

Amplicon sequencing

Amplicon DNA was prepared using the Illumina COVIDSeq DNA prep kit with primer pools for either *S. pneumoniae* or *M. tuberculosis* alongside a negative template control as performed previously [65]. Template DNA from each specimen was amplified in two separate PCRs, one reaction for each primer pool. For each sample, equal amounts of each PCR product were combined and the 2kb amplification products underwent tagmentation for 3 minutes followed by a bead cleanup and library amplification with Illumina index adapters. Equal volumes of the fragmented and indexed library for each sample was pooled, followed by size-selective bead cleanup for DNA fragments between 300-600 bp. The final pooled library was quantified with a Qubit fluorometer and dsDNA High-Sensitivity Assay kit, and the fragment distribution verified on an Agilent Bioanalyzer and high-sensitivity DNA kit. Pooled libraries were sequenced on an Illumina NovaSeq (paired-end 150) with an average of 10 million reads per library.

Alignments & Calling

Reads were aligned to the appropriate reference (*S. pneumoniae*: CC180 (Serotype 3); *M. tuberculosis:* H37Rv) using BWA-MEM (v.2.2.1) [66] and SAMtools (v1.15.1) [67]. Amplicon sequencing data were filtered (using defaults; Q>20 over a sliding window of 4, minimum read length 50% of the average length). TB primer sequences were trimmed using iVar (v.1.4.2) [68]. Metagenomic sequences were trimmed and filtered for quality and length (<100bp), using Trim Galore (v.0.6.10) [69]. Variants were called and filtered (Phred score Q>10 and read depth >10) using BCFtools [70]. Read subsampling, depth, and coverage was calculated using SAMtools [67]. Raw reads were directly submitted to the CZID mNGS Illumina pipeline [71] for microbial composition characterization within samples. Further data analyses and visualizations were carried out in Rstudio (v.2024.04.2+764) [72] using the tidyverse suite (v.2.0.0) [73].

Off-target amplification prediction

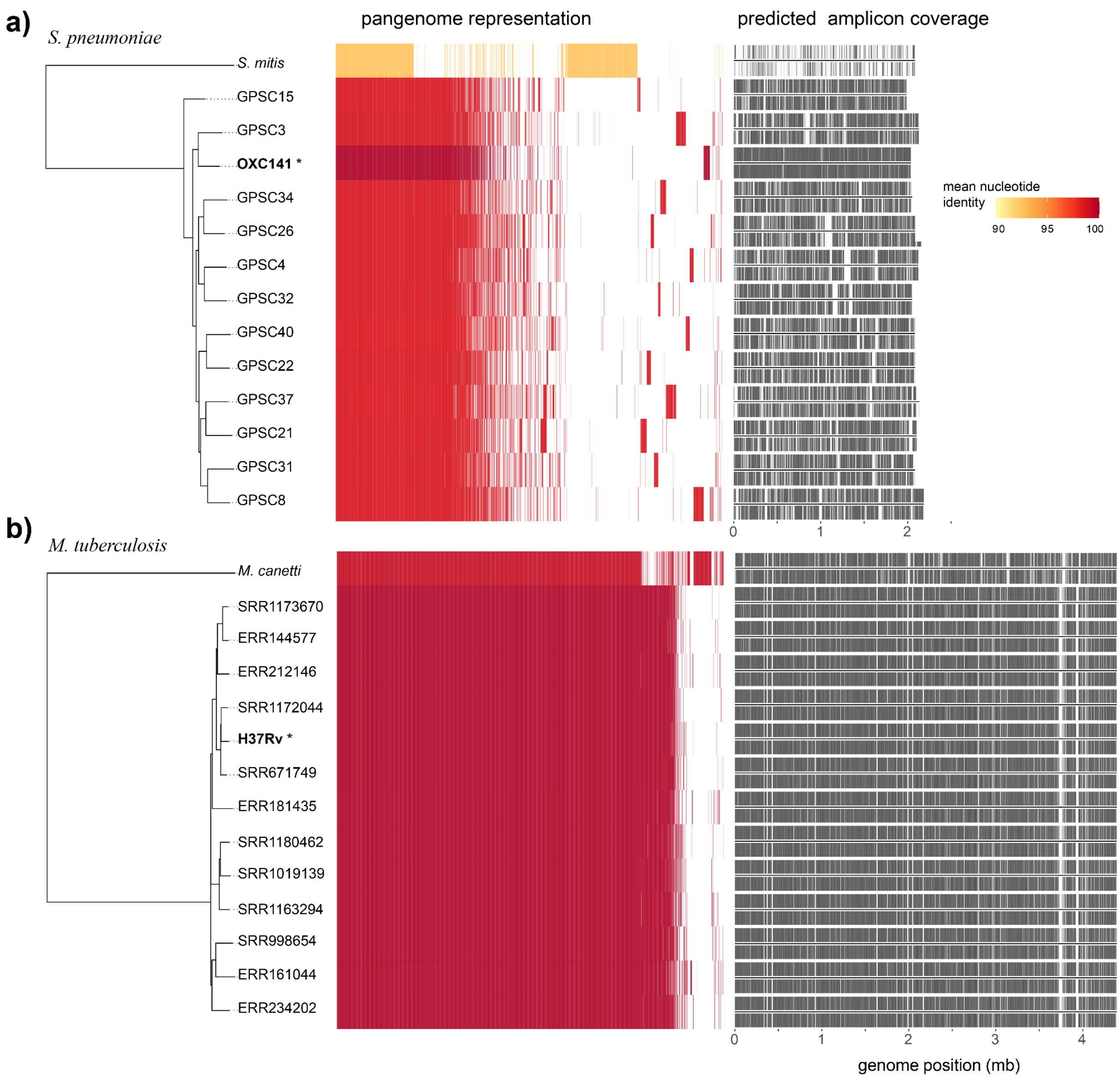
For each amplicon panel, off-target amplification was assessed *in-silico* against a set of related genomes. For each species we compiled a genome cluster consisting of the reference genome, 12 representative near-neighbor genomes, and an outgroup (**Supplemental Table 2**). The pangenome for each cluster was calculated using Roary (v.3.13.0) [74] and assembled a maximum-likelihood (ML) phylogeny using FastTree (v.2.1.11) [75]. Average nucleotide distance was calculated between out references and all other genomes in the cluster using FastANI (v.1.34) [76]. Off-target amplification was inferred by primer alignment using Bowtie (v.1.3.1) [77]; amplicons were predicted for any properly oriented amplicon pairs within 2,200 bp.

Serotyping, lineage assignment, and resistance prediction

S. pneumoniae isolates were *de novo* assembled with Shovill (v.1.1.0) [78]. *In-silico* multi-locus sequencing types (MLST) were assigned with mlst (v.2.23.0) [79]. Global pneumococcal sequencing clusters (GPSCs) were assigned with poppunk (v.2.7.0) [24]. *In-silico* screening of contigs for *S. pneumoniae* antimicrobial and virulence genes was done using ABRicate (v1.0.1) [80] and appropriate AMR databases [81–84]. For M. tuberculosis, Mykrobe [21] was used to both assign lineages and predict resistance using the built-in panel “202309” for tuberculosis [85]. As a comparison, a time-resolved maximum-likelihood tree was built using our variant call file, along with sample data generated from Augur (v.22.4.0), IQ-Tree (v.2.23), and TreeTime (v.0.10.1). Tree visualisations were done using Auspice (v2.57.0).

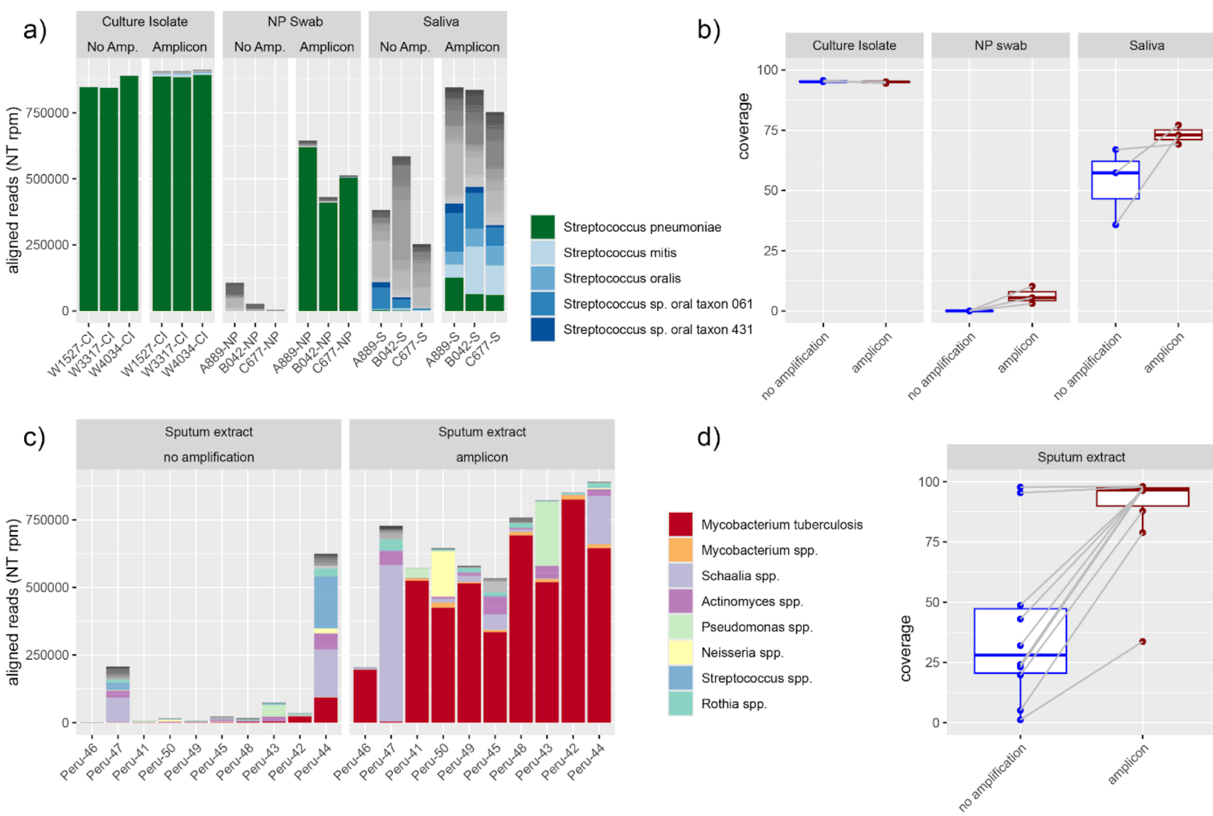
Figures

Figure 1: *In silico* modeling indicates broad applicability across diverse TB clades



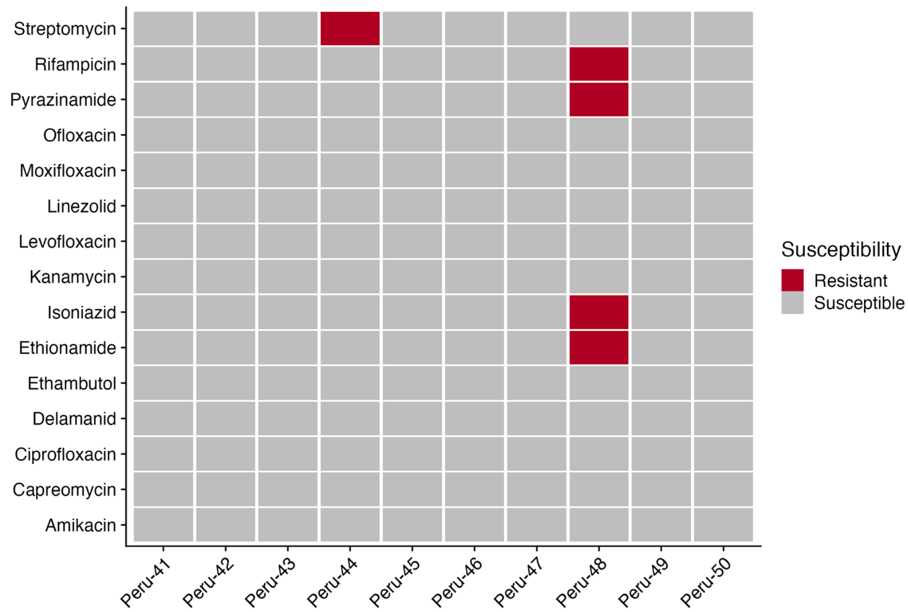
Pangenome representation of (A) *S. pneumoniae* whole genome sequences (n=13) and *S. mitis* outgroup (Accession: AP023349) and (B) *M. tuberculosis* whole genome sequences (n=13) and *M. canetti* outgroup (Accession: NC_019950). Starred phylogenetic tree tips mark the reference sequences used for primer design. Shaded bar graphs (middle) denote genes shared amongst clades, color denotes average nucleotide identity. Predicted amplicon coverage (right) is shown in grey with forward and reverse amplicon pairs displayed above and below the line. A table of the sequences used in this analysis can be found at **Supplemental Table 2**.

Figure 2: Tiled amplicon sequencing enables recovery of whole genome sequences from TB sputum



Comparisons between amplified and unamplified clinical samples were made for both species with regard to metagenomics (a,c), via the CZID metagenomics pipeline, and overall genome coverage (b,d). *S. pneumoniae* samples from multiple sample types from matched patients (a-b) showed increases in genome coverage and depth for all sample types, despite simultaneous amplification of closely related taxa. *M. tuberculosis* samples taken from direct sputum sequencing (c-d) show dramatic increases in genome coverage, with 8/10 samples generating more than 80% coverage after amplification with our protocol, and a ninth sample generating 78% coverage despite a significant infection with *Schaalia odontolytica*.

Figure 3: Amplicon sequencing predicts TB antimicrobial resistance *in-silico*.



Predicted susceptibility to 15 anti-TB drugs by amplicon sequencing for DNA extracted from sputum without prior culture using our optimised extraction protocol, showing detection of Streptomycin and combined Rifampicin / Isoniazid resistance.

Funding statement

This publication was made possible by the New England Pathogen Genomics Center of Excellence (US CDC NU50CK000629); The National Heart, Lung, and Blood Institute of the National Institutes of Health and the Richard K. Gershon Endowed Medical Student Fellowship at Yale University School of Medicine.

Citations

1. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat Protoc. 2017;12: 1261–1276.

2. Artic Network. [cited 30 Aug 2024]. Available: https://artic.network/ncov-2019

3. Hill V, Githinji G, Vogels CBF, Bento AI, Chaguza C, Carrington CVF, et al. Toward a global virus genomic surveillance network. Cell Host Microbe. 2023;31: 861–873.

4. Langedijk AC, Lebbink RJ, Naaktgeboren C, Evers A, Viveen MC, Greenough A, et al. Global molecular diversity of RSV - the "INFORM RSV" study. BMC Infect Dis. 2020;20: 450.

5. Tulloch RL, Kok J, Carter I, Dwyer DE, Eden J-S. An Amplicon-Based Approach for the Whole-Genome Sequencing of Human Metapneumovirus. Viruses. 2021;13. doi:10.3390/v13030499

6. Vogels CBF, Hill V, Breban MI, Chaguza C, Paul LM, Sodeinde A, et al. DengueSeq: a pan-serotype whole genome amplicon sequencing protocol for dengue virus. BMC Genomics. 2024;25: 433.

7. Chen NFG, Chaguza C, Gagne L, Doucette M, Smole S, Buzby E, et al. Development of an amplicon-based sequencing approach in response to the global emergence of mpox. PLoS Biol. 2023;21: e3002151.

8. Ikuta KS, Swetschinski LR, Robles Aguilar G, Sharara F, Mestrovic T, Gray AP, et al. Global mortality associated with 33 bacterial pathogens in 2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet. 2022;400: 2221–2248.

9. Djordjevic SP, Jarocki VM, Seemann T, Cummins ML, Watt AE, Drigo B, et al. Genomic surveillance for antimicrobial resistance - a One Health perspective. Nat Rev Genet. 2024;25: 142–157.

10. Chen Y, Luo Y, Carleton H, Timme R, Melka D, Muruvanda T, et al. Whole Genome and Core Genome Multilocus Sequence Typing and Single Nucleotide Polymorphism Analyses of Listeria monocytogenes Isolates Associated with an Outbreak Linked to Cheese, United States, 2013. Appl Environ Microbiol. 2017;83. doi:10.1128/AEM.00633-17

11. Dallman T, Inns T, Jombart T, Ashton P, Loman N, Chatt C, et al. Phylogenetic structure of European Salmonella Enteritidis outbreak correlates with national and international egg distribution network. Microb Genom. 2016;2: e000070.

12. Li Z, Pérez-Osorio A, Wang Y, Eckmann K, Glover WA, Allard MW, et al. Whole genome sequencing analyses of Listeria monocytogenes that persisted in a milkshake machine for a year and caused illnesses in Washington State. BMC Microbiol. 2017;17: 134.

13. Luterbach CL, Chen L, Komarow L, Ostrowsky B, Kaye KS, Hanson B, et al. Transmission of Carbapenem-Resistant Klebsiella pneumoniae in US Hospitals. Clin Infect Dis. 2023;76: 229–237.

14. Snitkin ES, Won S, Pirani A, Lapp Z, Weinstein RA, Lolans K, et al. Integrated genomic and interfacility patient-transfer data reveal the transmission pathways of multidrug-resistant *Klebsiella pneumoniae* in a regional outbreak. Sci Transl Med. 2017;9. doi:10.1126/scitranslmed.aan0093

15. Tong SYC, Holden MTG, Nickerson EK, Cooper BS, Köser CU, Cori A, et al. Genome sequencing defines phylogeny and spread of methicillin-resistant Staphylococcus aureus in a high transmission setting. Genome Res. 2015;25: 111–118.

16. Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, et al. Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular

epidemiological study. PLoS Med. 2013;10: e1001387.

17. Lee RS, Proulx J-F, McIntosh F, Behr MA, Hanage WP. Previously undetected super-spreading of Mycobacterium tuberculosis revealed by deep sequencing. Elife. 2020;9. doi:10.7554/eLife.53245

18. Hatherell H-A, Didelot X, Pollock SL, Tang P, Crisan A, Johnston JC, et al. Declaring a tuberculosis outbreak over with genomic epidemiology. Microbial Genomics. 2016;1: 10.1099/mgen.0.000060.

19. Pankhurst LJ, Del Ojo Elias C, Votintseva AA, Walker TM, Cole K, Davies J, et al. Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. Lancet Respir Med. 2016;4: 49–58.

20. Cohen KA, Manson AL, Desjardins CA, Abeel T, Earl AM. Deciphering drug resistance in Mycobacterium tuberculosis using whole-genome sequencing: progress, promise, and challenges. Genome Med. 2019;11: 45.

21. Hunt M, Bradley P, Lapierre SG, Heys S, Thomsit M, Hall MB, et al. Antibiotic resistance prediction for Mycobacterium tuberculosis from genome sequence data with Mykrobe. Wellcome Open Res. 2019;4: 191.

22. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37: 1530–1534.

23. Sheppard CL, Manna S, Groves N, Litt DJ, Amin-Chowdhury Z, Bertran M, et al. PneumoKITy: A fast, flexible, specific, and sensitive tool for Streptococcus pneumoniae serotype screening and mixed serotype detection from genome sequence data. Microb Genom. 2022;8. doi:10.1099/mgen.0.000904

24. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Res. 2019;29: 304–316.

25. Inouye M, Dashnow H, Raven L, Schultz MB, Pope BJ, Tomita T, et al. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. bioRxiv. bioRxiv; 2014. doi:10.1101/006627

26. Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet. 2018;19: 9–20.

27. Bentley SD, Lo SW. Global genomic pathogen surveillance to inform vaccine strategies: a decade-long expedition in pneumococcal genomics. Genome Med. 2021;13: 84.

28. Ledesma JR, Ma J, Vongpradith A, Maddison ER, Novotney A, Biehl MH, et al. Global, regional, and national sex differences in the global burden of tuberculosis by HIV status, 1990–2019: results from the Global Burden of Disease Study 2019. Lancet Infect Dis. 2022;22: 222–241.

29. Kunii O, Yassin MA, Wandwalo E. Investing to end epidemics: the role of the Global Fund to control TB by 2030. Trans R Soc Trop Med Hyg. 2016;110: 153–154.

30. World Health Organization. Global tuberculosis report 2023. World Health Organization; 2023.

31. WHO consolidated guidelines on tuberculosis: Module 4: treatment - drug-resistant tuberculosis treatment, 2022 update. Geneva: World Health Organization; 2022.

32. Tseng S-Y, Huang Y-S, Chang T-E, Perng C-L, Huang Y-H. Hepatotoxicity, efficacy and completion rate between 3 months of isoniazid plus rifapentine and 9 months of isoniazid in treating latent tuberculosis infection: A systematic review and meta-analysis: A systematic review and meta-analysis. J Chin Med Assoc. 2021;84: 993–1000.

33. World Health Organization. WHO consolidated guidelines on tuberculosis. Module 3: diagnosis – rapid diagnostics for tuberculosis detection. World Health Organization; 2024.

34. Walker TM, Miotto P, Köser CU, Fowler PW, Knaggs J, Iqbal Z, et al. The 2021 WHO catalogue of Mycobacterium tuberculosis complex mutations associated with drug resistance: A genotypic analysis. Lancet Microbe. 2022;3: e265–e273.

35. Kadura S, King N, Nakhoul M, Zhu H, Theron G, Köser CU, et al. Systematic review of mutations associated with resistance to the new and repurposed Mycobacterium tuberculosis drugs bedaquiline, clofazimine, linezolid, delamanid and pretomanid. J Antimicrob Chemother. 2020;75: 2031–2043.

36. Doughty EL, Sergeant MJ, Adetifa I, Antonio M, Pallen MJ. Culture-independent detection and characterisation of Mycobacterium tuberculosis and M. africanum in sputum samples using shotgun metagenomics on a benchtop sequencer. PeerJ. 2014;2: e585.

37. Votintseva AA, Bradley P, Pankhurst L, Del Ojo Elias C, Loose M, Nilgiriwala K, et al. Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples. J Clin Microbiol. 2017;55: 1285–1298.

38. Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZM, et al. Rapid Whole-Genome Sequencing of Mycobacterium tuberculosis Isolates Directly from Clinical Samples. J Clin Microbiol. 2015;53: 2230–2237.

39. McNerney R, Clark TG, Campino S, Rodrigues C, Dolinger D, Smith L, et al. Removing the bottleneck in whole genome sequencing of Mycobacterium tuberculosis for rapid drug resistance analysis: a call to action. Int J Infect Dis. 2017;56: 130–135.

40. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. The Lancet Infectious diseases. 2013;13. doi:10.1016/S1473-3099(12)70277-3

41. Bryant JM, Schürch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V, et al. Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data. BMC Infect Dis. 2013;13: 110.

42. Yang C, Sobkowiak B, Naidu V, Codreanu A, Ciobanu N, Gunasekera KS, et al. Phylogeography and transmission of M. tuberculosis in Moldova: A prospective genomic analysis. PLoS Med. 2022;19: e1003933.

43. Guerra-Assunção JA, Crampin AC, Houben RMGJ, Mzembe T, Mallard K, Coll F, et al. Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. Elife. 2015;4. doi:10.7554/eLife.05166

44. Qian J, Chen R, Wang H, Zhang X. Role of the PE/PPE family in host-pathogen interactions and prospects for anti-tuberculosis vaccine and diagnostic tool design. Front Cell Infect Microbiol. 2020;10: 594288.

45. D'Souza C, Kishore U, Tsolaki AG. The PE-PPE Family of Mycobacterium tuberculosis: Proteins in Disguise. Immunobiology. 2023;228: 152321.

46. Sadowy E, Bojarska A, Kuch A, Skoczyńska A, Jolley KA, Maiden MCJ, et al. Relationships among streptococci from the mitis group, misidentified as Streptococcus pneumoniae. Eur J Clin Microbiol Infect Dis. 2020;39: 1865–1878.

47. Kalizang'oma A, Chaguza C, Gori A, Davison C, Beleza S, Antonio M, et al. Streptococcus pneumoniae serotypes that frequently colonise the human nasopharynx are common recipients of penicillin-binding protein gene fragments from Streptococcus mitis. Microb Genom. 2021;7. doi:10.1099/mgen.0.000622

48. Hunt M, Lima L, Shen W, Lees J, Iqbal Z. AllTheBacteria - all bacterial genomes assembled, available and searchable. bioRxiv. 2024. p. 2024.03.08.584059. doi:10.1101/2024.03.08.584059

49. Bobay L-M, Ochman H. Factors driving effective population size and pan-genome evolution in bacteria. BMC Evol Biol. 2018;18: 153.

50. Manson AL, Cohen KA, Abeel T, Desjardins CA, Armstrong DT, Barry CE 3rd, et al. Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. Nat Genet. 2017;49: 395–402.

51. Ntagereka PB, Oyola SO, Baenyi SP, Rono GK, Birindwa AB, Shukuru DW, et al. Whole-genome sequencing of SARS-CoV-2 reveals diverse mutations in circulating Alpha and Delta variants during the first, second, and third waves of COVID-19 in South Kivu, east of the Democratic Republic of the Congo. Int J Infect Dis. 2022;122: 136–143.

52. Khairnar K, Tomar SS. COVID-19 genome surveillance: A geographical landscape and mutational mapping of SARS-CoV-2 variants in central India over two years. Virus Res. 2024;344: 199365.

53. Cowley LA, Afrad MH, Rahman SIA, Mamun MMA, Chin T, Mahmud A, et al. Genomics, social media and mobile phone data enable mapping of SARS-CoV-2 lineages to inform health policy in Bangladesh. Nat Microbiol. 2021;6: 1271–1278.

54. Saha S, Tanmoy AM, Hooda Y, Tanni AA, Goswami S, Sium SMA, et al. COVID-19 rise in Bangladesh correlates with increasing detection of B.1.351 variant. BMJ Glob Health. 2021;6: e006012.

55. Jony MHK, Alam AN, Nasif MAO, Sultana S, Anwar R, Rudra M, et al. Emergence of SARS-CoV-2 Omicron sub-lineage JN.1 in Bangladesh. Microbiol Resour Announc. 2024;13: e0013024.

56. Denkinger CM, Schumacher SG, Gilpin C, Korobitsyn A, Wells WA, Pai M, et al. Guidance for the evaluation of tuberculosis diagnostics that meet the World Health Organization (WHO) target product profiles: An introduction to WHO process and study design principles. J Infect Dis. 2019;220: S91–S98.

57. Koo H-K, Min J, Kim HW, Lee J, Kim JS, Park JS, et al. Prediction of treatment failure and compliance in patients with tuberculosis. BMC Infect Dis. 2020;20: 622.

58. Liu Q, Zhu J, Dulberger CL, Stanley S, Wilson S, Chung ES, et al. Tuberculosis treatment failure associated with evolution of antibiotic resilience. Science. 2022;378: 1111–1118.

59. van der Werf MJ, Langendam MW, Huitric E, Manissero D. Multidrug resistance after inappropriate tuberculosis treatment: a meta-analysis. Eur Respir J. 2012;39: 1511–1519.

60. Gladstone RA, Lo SW, Lees JA, Croucher NJ, van Tonder AJ, Corander J, et al. International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. EBioMedicine. 2019;43: 338–346.

61. Carvalho M da GS, Tondella ML, McCaustland K, Weidlich L, McGee L, Mayer LW, et al. Evaluation and improvement of real-time PCR assays targeting lytA, ply, and psaA genes for detection of pneumococcal DNA. J Clin Microbiol. 2007;45: 2460–2466.

62. Trzciński K, Bogaert D, Wyllie A, Chu MLJN, van der Ende A, Bruin JP, et al. Superiority of trans-oral over trans-nasal sampling in detecting Streptococcus pneumoniae colonization in adults. PLoS One. 2013;8: e60520.

63. Wyllie AL, Mbodj S, Thammavongsa DA, Hislop MS, Yolda-Carr D, Waghela P, et al. Persistence of Pneumococcal Carriage among Older Adults in the Community despite COVID-19 Mitigation Measures. Microbiol Spectr. 2023;11: e0487922.

64. Goig GA, Torres-Puente M, Mariner-Llicer C, Villamayor LM, Chiner-Oms Á, Gil-Brusola A, et al. Towards next-generation diagnostics for tuberculosis: identification of novel molecular targets by large-scale comparative genomics. Bioinformatics. 2020;36: 985–989.

65. Chen NFG, Gagne L, Doucette M, Smole S, Buzby E, Hall J, et al. Monkeypox virus multiplexed PCR amplicon sequencing (PrimalSeq). protocols.io. 2022 [cited 26 Aug 2024]. doi:10.17504/protocols.io.5qpvob1nbl4o/v4

66. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-bio.GN]. 2013. Available: http://arxiv.org/abs/1303.3997

67. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.

68. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol. 2019;20: 8.

69. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17: 10–12.

70. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10. doi:10.1093/gigascience/giab008

71. Lu D, Kalantar KL, Chu VT, Glascock AL, Guerrero ES, Bernick N, et al. Simultaneous detection of pathogens and antimicrobial resistance genes with the open source, cloud-based, CZ ID pipeline. bioRxivorg. 2024 [cited 30 Aug 2024]. doi:10.1101/2024.04.12.589250

72. Allaire J. RStudio: integrated development environment for R. Boston, MA. 2012;770: 165–171.

73. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the tidyverse. J Open Source Softw. 2019;4: 1686.

74. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31: 3691–3693.

75. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009;26: 1641–1650.

76. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9: 5114.

77. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10: R25.

78. Seemann T, Others. Shovill: Faster SPAdes assembly of Illumina reads. 2017. 2022.

79. Seemann T. mlst: :id: Scan contig files against PubMLST typing schemes. Github; Available: https://github.com/tseemann/mlst

80. Seemann T. abricate: :mag_right: Mass screening of contigs for antimicrobial and virulence genes. Github; Available: https://github.com/tseemann/abricate

81. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res. 2017;45: D566–D573.

82. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother. 2014;58: 212–220.

83. Chen L, Zheng D, Liu B, Yang J, Jin Q. VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on. Nucleic Acids Res. 2016;44: D694–7.

84. Doster E, Lakin SM, Dean CJ, Wolfe C, Young JG, Boucher C, et al. MEGARes 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data.

Nucleic Acids Res. 2020;48: D561–D569.

85. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil P-A, Hugenholtz P. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. 2022;50: D785–D794.