Nimmo, Camus; Shaw, Liam; Doyle, Ronan; Williams, Rachel; Brien, Kayleen; Burgess, Carrie; Breuer, Judith; Balloux, Francois; Pym, Alexander; (2018) Whole genome sequencing Mycobacterium tuberculosis directly from sputum identifies more genetic diversity than sequencing from culture. DOI: https://doi.org/10.1101/446849

Downloaded from: http://researchonline.lshtm.ac.uk/4652518/

DOI: https://doi.org/10.1101/446849

Usage Guidelines:

Please refer to usage guidelines at https://researchonline.lshtm.ac.uk/policies.html or alternatively contact researchonline@lshtm.ac.uk.

Available under license: http://creativecommons.org/licenses/by-nc-nd/2.5/
Article type: Research

Title: Whole genome sequencing *Mycobacterium tuberculosis* directly from sputum identifies more genetic diversity than sequencing from culture

Authors:
Camus Nimmo\(^1,2\) c.nimmo.04@cantab.net
Liam P. Shaw\(^3,4\)
Ronan Doyle\(^1\)
Rachel Williams\(^1\)
Kayleen Brien\(^2\)
Carrie Burgess\(^1\)
Judith Breuer\(^1\)
Francois Balloux\(^3\)
Alexander S. Pym\(^2\)

1. Division of Infection and Immunity, University College London, London WC1E 6BT, UK
2. Africa Health Research Institute, Durban, South Africa
3. UCL Genetics Institute, University College London, London WC1E 6BT, UK
4. Nuffield Department of Clinical Medicine, Oxford University, Oxford OX3 7BN, UK
Abstract

Background
Repeatecultu re reduces within-sample Mycobacterium tuberculosis genetic diversity due to selection of clones suited to growth in culture and/or random loss of lineages, but it is not known to what extent omitting the culture step altogether alters genetic diversity. We compared M. tuberculosis whole genome sequences generated from 39 paired clinical samples. In one sample DNA was extracted directly from sputum then enriched with custom-designed SureSelect (Agilent) oligonucleotide baits and in the other it was extracted from mycobacterial growth indicator tube culture.

Results
DNA directly sequenced from sputum showed significantly more within-sample diversity than that from mycobacterial growth incubator tube culture. This was demonstrated by more variants present as heterozygous alleles (HAs) where both a variant and wild type allele were present within a given sample (p<0.001) and greater within-sample Shannon diversity (p<0.001). Seven genes with high within-sample diversity have previously been identified as targets for positive selection, highlighting their potential role in adaptation to survival within the host and under drug pressure. Resistance associated variants present as HAs occurred in six patients, and in four cases may have provided a genotypic explanation for phenotypic resistance.

Conclusions
Culture-free *M. tuberculosis* whole genome sequencing detects more within-sample diversity and may allow detection of mycobacteria that are not actively replicating.

Key words: Mycobacterium tuberculosis; drug-resistant tuberculosis; whole genome sequencing; sputum; within-patient diversity; heteroresistance
**Background**

International efforts to reduce tuberculosis (TB) infections and mortality over the last two decades have only been partially successful. In 2017, 10 million people developed TB and it has overtaken HIV as the infectious disease responsible for the most deaths worldwide(1, 2). Drug resistance is a major concern with a steady rise in the number of reported cases globally and rapid increases in some areas(1). Patients with *Mycobacterium tuberculosis* resistant to the first line drugs rifampicin and isoniazid are classed as having multidrug-resistant (MDR) TB and usually treated with a standardised second-line drug regimen for at least nine months, which is also used for rifampicin monoresistance(3, 4). With the emergence of resistance to fluoroquinolones and aminoglycosides (extensively drug-resistant [XDR] TB) there is an increasing need for individualised therapy based on drug susceptibility testing (DST). Individualised therapy ensures patients are treated with sufficient active drugs which can prevent selection of additional resistance, improve treatment outcomes and reduce duration of infectiousness(5-8).

Traditionally, phenotypic culture-based DST was used to identify drug resistance but this is being replaced by rapid genetic tests that detect specific drug resistance conferring mutations. Next generation whole genome sequencing (WGS) of *M. tuberculosis* is being increasingly used in research and clinical settings to comprehensively identify all drug resistance associated mutations(9). *M. tuberculosis* has a conserved genome with little genetic diversity between strains(10), but more detailed analysis of individual patient samples with WGS has identified genetically separate bacterial subpopulations in sequential sputum samples(11-15) and across different anatomical sites(16). This within-patient
diversity can occur as a result of mixed infection with genetically distinct strains or within-host evolution of a single infecting strain (17).

Bacterial subpopulations can be detected in clinical samples after sequencing reads are mapped to a reference genome where multiple base calls are detected at a single genomic site. These heterozygous alleles (HAs) at sites associated with drug resistance (resistance associated variants, RAVs) may reflect heteroresistance, where a fraction of the total bacterial population is drug susceptible while the remainder is resistant (18). Identification of genetic diversity within clinical samples is important as it may improve detection of RAVs over currently available genetic tests and consensus-level WGS (18). Identifying RAVs could improve individualised therapy, prevent acquired resistance (12), and give insight into bacterial adaptation to the host.

*M. tuberculosis* WGS is usually performed on cultured isolates to obtain sufficient purified mycobacterial DNA. However, the culture process can change the population structure from that of the original sample due to genetic drift (random loss of lineages) and/or the selection of subpopulations more suited to growth in culture (19-21), and repeated subculture leads to loss of genetic diversity and heteroresistance (22). Additionally, in the normal course of *M. tuberculosis* infection, some bacteria exist as viable non-culturable persister organisms that are hypothesised to cause the high relapse rate seen following treatment of insufficient duration (23). These organisms are likely to be missed by any sequencing method reliant on culture.
WGS directly from sputum without enrichment is challenging (24). It has recently been improved by depleting human DNA during DNA extraction (25). We have previously reported the use of oligonucleotide enrichment technology SureSelect (Agilent, CA, USA) to sequence *M. tuberculosis* DNA directly from sputum (26) and demonstrated its utility in determining a rapid genetic drug resistance profile (27, 28).

It remains unclear to what extent WGS of cultured *M. tuberculosis* samples underestimates the genetic diversity of the population in sputum samples. One previous study of 16 patients did not identify increased genetic diversity in *M. tuberculosis* DNA sequenced directly from sputum compared to DNA from culture (25), whereas another study of mostly drug susceptible patients showed sequencing directly from sputum identified a slight excess of HAs relative to culture (27). Here we reanalyse heterozygous alleles (HAs) present in that study (27) in addition to newly collected samples from patients with MDR-TB, use a more sensitive analysis to measure overall within-sample genetic diversity and further explore the genomic location of the additional diversity identified.

**Results**

**Patient Characteristics and Drug Susceptibility Testing**

Whole genome sequences were obtained for 39 patients from both mycobacterial growth indicator tube (MGIT) culture and direct sputum sequencing. The patients were predominantly of black African ethnicity (80%) and 50% were HIV positive (Table 1). First-line phenotypic drug susceptibility testing (DST) results identified 22 patients with MDR-TB
and two with rifampicin monoresistance. In addition there were three isoniazid
monoresistant patients and ethambutol resistance was detected in 8 patients. Second-line
phenotypic DST was performed for patients with rifampicin-resistant or MDR-TB and
identified one case of kanamycin resistance (Table 2).

We observed greater median coverage depth in sputum-derived sequences than MGIT
sequences (164.3 vs 136.6, p=0.068). A genotypic susceptibility profile was determined by
evaluating MGIT WGS for consensus-level RAVs using a modified version of a publicly
available list(29). Genotypic RAVs predicted all rifampicin phenotypic resistance and >90% of
isoniazid phenotypic resistance. Ethambutol genotypic RAVs were poorly predictive of
phenotypic resistance in line with findings from other studies(30) (Table 2). The patient with
kanamycin phenotypic resistance was correctly identified by an rrs a1401g RAV. No full
phenotypic fluoroquinolone phenotypic resistance was identified, but several colonies from
patient F1013 did grow in the presence of ofloxacin (although not enough to be classified as
resistant). The consensus sequences from this patient harboured a gyrB E501D mutation
which is believed to confer resistance to moxifloxacin but not other fluoroquinolones, which
may explain the borderline phenotypic DST result(31).

Genetic Diversity

To compare consensus sequences from sputum and MGIT, a WGS consensus sequence-level
maximum likelihood phylogenetic tree was constructed (Supplementary Material: Figure 1).
Four previously sequenced strains from KwaZulu-Natal were included(32). As expected, all
paired sequences were closely related, with a mean difference of 1.30 (range 0-9) single
nucleotide polymorphisms (SNPs). Samples from patients F1066 and F1067 were closely related with only one consensus-level SNP separating all four consensus sequences. There was no obvious epidemiological link between these patients (although this study was not designed to collect comprehensive epidemiological information) and they lived 20km apart in Durban. However, both patients were admitted contemporaneously to an MDR treatment facility and sampled on the same day. DNA extraction and sequencing occurred on different runs so the close genetic linkage may represent direct transmission within a hospital setting, a community transmission chain or an unlikely cross-contamination during sample collection.

Having established congruence between sputum and MGIT sequences at the consensus level we then compared genetic diversity by DNA source. We first defined a threshold for calling variants present as heterozygous alleles (HAs) in our entire dataset by using a range of minimum read count frequencies as described in the methods (Figure 1). Below a minimum of five supporting reads there was an exponential increase in the number of HAs identified, which may be indicative of the inclusion of sequencing errors. To reduce this risk, we used a threshold of a minimum of five supporting reads.

Genetic diversity may occur because of within-host evolution or mixed infection. To identify mixed infection we used a molecular barcode(33) to scan all HAs for a panel of 413 phylogenetic SNPs that can resolve *M. tuberculosis* into one of seven lineages and 55 sublineages. We found three phylogenetic SNPs among the HAs. In all cases the heterozygous phylogenetic SNP originated from the same sublineage as other SNPs present at 100% frequency, and there were no cases of HAs indicating the presence of more than one lineage
or sublineage. This suggests that the genetic diversity identified is mostly or exclusively due
to within-host evolution, although there remains a small possibility that mixed infections
with two strains from the same sub-lineage could have occurred.

As a first step to comparing diversity between sputum and MGIT sequenced samples we
looked at the location of genetic diversity within the \textit{M. tuberculosis} genome. Variants were
called in the MGIT and sputum sequences for each patient and classified as present in MGIT
only, sputum only or shared (present in both). HAs were widely dispersed across the
genome at similar sites in both sputum and MGIT samples but some genes had multiple HAs
(Table 3). The highest genetic diversity was found in the ribosomal RNA (rRNA) genes \((rrs
\text{ and } rrl)\) with 358 HAs, of which 98.6\% were only found in sputum-derived sequences.

As rRNA contains regions that are highly conserved across bacteria, it was considered a
possibility that SureSelect baits targeting rRNA genes were capturing both \textit{M. tuberculosis}
and other bacterial species. To evaluate this, metagenomic assignment was performed on
all reads. Sampling reads not assigned to \textit{M. tuberculosis} (i.e. presumed contaminants from
other bacteria) and performing a BLAST search against \textit{M. tuberculosis} 16S and 23S rRNA
genes indicated that a sizeable proportion of these reads from directly sequenced sputum
had a BLAST hit of at least 30 bases (median 11\% vs 0\% of equivalent reads from MGIT
sequencing, \(p<0.001\), Supplementary Material: Figure 2). The taxonomic assignment of
these reads were indeed typical of genera composing the oral flora, with a high
representation of \textit{Actinomyces}, \textit{Fusobacterium}, \textit{Prevotella}, and \textit{Streptococcus}
(Supplementary Material: Figure 3).
This supported the hypothesis that the baits may enrich rRNA from other organisms so rRNA genes were excluded from further analysis. The difference in diversity between sputum and MGIT sequences can be explained by the selective nature of MGIT media which will enrich *M. tuberculosis* sequences. Importantly the frequency of HAs in other highly diverse genes between sequencing strategies was more balanced (Table 3). Pertinently seven of these genes (*Rv1319c* (34), *lppB* (35), *Rv2082* (35), *ppsA* (34, 36), *ponA* (36), *lppA* (37), and *pkS* (35, 36)) with high numbers of HAs have been previously identified as highly diverse in comparative genomic studies suggesting the detected HAs are not artefactual. The frequencies at which HAs in these genes were present in MGIT and sputum is shown in Supplementary Material Figure 4.

After confirming the absence of mixed infections and removing rRNA gene sequences we compared the frequency of HAs in sputum and MGIT. There were 2048 variants across the dataset that were present as a HA in either MGIT, sputum or both sequences (Table 4). Variants present in both MGIT and sputum derived sequences were more likely to be present as a HA in the sputum-derived sequence (3.2% v 1.9%, p<0.0001). Of the other variants present as HAs, 821 were unique to direct sputum sequencing and 153 were unique to MGIT sequencing (Table 4). Variants found only in sputum were more likely to be heterozygous than those found only in MGIT or in both (p<0.0001). The distribution of HAs by patient is shown in Figure 2A. HAs found only by one modality were more likely to be SNPs than shared HAs, where the majority were insertions or deletions. The ratio of non-synonymous to synonymous HAs was similar for those that were shared or MGIT only, but was lower for sputum only HAs. Frameshift mutations were most prevalent among shared HAs (Table 4).
To confirm our findings of increased diversity in sputum we calculated mean within-sample diversity ($H$), excluding rRNA genes and repeat regions (see methods). The mean diversity was significantly greater in sputum than MGI T-derived sequences (Figure 2B: $0.116 \pm 0.078$ v $0.054 \pm 0.026$, mean $H_{\text{sputum}}/H_{\text{MGIT}} = 2.66$, $p=3.0 \times 10^{-5}$).

Genetic diversity in drug resistance genes

HAs in drug resistance-associated regions, including promoters and intergenic regions, were individually assessed. Five of the 39 patients had RAVs present as HAs in at least one gene, which are shown in Table 5. F1002 had three compensatory mutations in $rpoC$ present at HAs in both sequences. F1007 had high-level phenotypic isoniazid resistance despite wild type $katG$ and $inhA$ genes, but did have two $ahpC$ promoter variants present as HAs. Neither of these variants are reported frequently but both have been previously associated with resistance in limited numbers of samples(38). As described above F1066 and F1067 were highly related with only one consensus SNP difference between all four sequences. Both had phenotypic high level isoniazid resistance with no consensus-level $katG$ or $inhA$ mutation, but had frameshift $katG$ mutations present as HAs which have the potential to cause resistance(39). F1066 and RF021 had $Rv1979c$ and $pncA$ mutations respectively at low frequency in sputum only which have the potential to confer phenotypic resistance to clofazimine ($Rv1979c$) and pyrazinamide ($pncA$), although no phenotypic testing was performed for these drugs.
In this study we whole genome sequenced DNA from sputum and MGIT culture in paired samples from 39 patients and compared within-patient genetic diversity of the bacterial genome identified from each source. All paired sequences were closely related at the consensus level, and WGS predicted phenotypic drug susceptibility with over 90% sensitivity and specificity for rifampicin and isoniazid in line with published data (40).

The understanding of within-patient *M. tuberculosis* genetic diversity is becoming increasingly important as the detection of rare variants has been shown to improve the correlation between phenotypic and genotypic drug resistance profiles (18) and can identify emerging drug resistance (11, 12). Here we have demonstrated that significantly more genetic diversity is identified by WGS performed directly from enriched sputum than MGIT culture. Not including a culture step avoids the introduction of bias towards culture-adapted subpopulations and the impact of random chance and is also likely to incorporate DNA from viable non-culturable mycobacteria. A reduction in genetic diversity has previously been shown with sequential *M. tuberculosis* subculture (19, 22), but was not confirmed by a study performing WGS directly from sputum (25). However, the 16 paired sputum and MGIT samples compared by Votintseva (25) had a minimum of 5x coverage compared to a minimum 40x coverage in this study, and were likely to contain less genetic material as they were surplus clinical rather than dedicated research samples.

We found that the rRNA genes have high levels of diversity in sputum samples, but believe this is due to non-mycobacterial DNA hybridising to the capture baits — a conclusion borne
out by the taxonomic assignment of reads aligning to these genes in common oral bacteria.

We therefore exclude these from further analysis, and recommend others using enrichment from sputum do similarly. We use two methods to evaluate within-sample \textit{M. tuberculosis} genetic diversity. First, we demonstrate increased diversity when sequencing directly from sputum with significantly more unique heterozygous alleles (HAs) than sequencing from MGIT culture. We also observed significantly higher genetic diversity in sputum-derived sequences by comparing the Shannon diversity of variable sites across pairs of samples.

Many of the genes with high levels of within-sample diversity are also reported to be targets for convergent evolution, independently accumulating similar mutations on a global scale. This implies that diversity seen on a macroevolutionary scale has a basis in microevolution, and reinforces the importance of accurately characterising the biological function of these genes and their products to aid the identification of new therapeutic targets. Two-thirds of the patients with MDR-TB had already been treated for drug-sensitive TB, and the diversity identified in sputum samples may therefore represent early adaptation to drug pressure.

Importantly, as direct sputum sequencing does not rely on live mycobacteria, DNA from recently killed \textit{M. tuberculosis} is likely to also be sequenced, meaning that recent genomic mutations are likely to be represented as HAs.

In four patients, RAVs present as HAs provided a likely genotypic basis for otherwise unexplained phenotypic resistance. Given the small total number of resistance mutations in this study, the excess of heterozygous known RAVs in directly sequenced sputum is not statistically significant. However the presence of heterozygous RAVs in both MGIT and sputum sequences reinforces the biological importance of these mutations.
A limitation of this study is that it can be difficult to distinguish low frequency variants from sequencing error. Ideally low frequency variants could be confirmed by resequencing the same DNA samples. To reduce the risk of sequencing errors yet still identify genetic diversity we used the lowest minimum read threshold at which the number of HAs remained stable. Also, it is reassuring that of all fixed and heterozygous variants called, more than 93% were identified from both DNA sources.

Conclusions

Directly sequencing *M. tuberculosis* from sputum is able to identify more genetic diversity than sequencing from culture. Understanding within-patient genetic diversity is important to understand bacterial adaptation to drug treatment and the acquisition of drug resistance. It also has potential to identify low frequency RAVs that may further enhance genotypic-phenotypic drug resistance correlation.

Methods

**Patient enrolment**

Adult patients presenting with a new diagnosis of sputum culture-positive TB were included in the study. Patients were recruited in London, UK (n=15) and Durban, South Africa (n=24). All patients recruited in Durban were Xpert MTB/RIF (Cepheid, CA, USA) positive for rifampicin resistance. Two sputum samples were collected prior to initiating treatment, with
one inoculated into mycobacterial growth indicator tube (MGIT) culture (BD, NJ, USA) and the other used for direct DNA extraction.

Ethics, Consent and Permissions

All patients gave written informed consent to participate in the study. Ethical approval for the London study was granted by NHS National Research Ethics Service East Midlands–Nottingham 1 (reference 15/EM/0091). Ethical approval for the Durban study was granted by University of KwaZulu-Natal Biomedical Research Ethics Committee (reference BE022/13).

Microbiology

MGIT samples were incubated in a BACTEC MGIT 960 (BD, NJ, USA) until flagging positive. Phenotypic DST data for London samples were those provided to treating hospitals by Public Health England. Phenotypic DST for Durban samples was performed using the solid agar proportion method (Supplementary Material: Methods).

DNA extraction and sequencing

Positive MGIT tubes were centrifuged at 16,000g for 15 minutes and the supernatant removed. Cells were resuspended in phosphate-buffered saline before undergoing heat killing at 95°C for 1 hour followed by centrifugation at 16,000g for 15 minutes. The supernatant was removed and the sample resuspended in 1mL sterile saline (0.9% w/v). The wash step was repeated. DNA was extracted with mechanical ribolysis before purification.
with DiaSorin Liaison Ixt (DiaSorin, Italy) or CTAB(41). NEBNext Ultra II DNA (New England Biolabs, MA, USA) was used for DNA library preparation.

Sputum samples for direct sequencing were similarly heat killed processed as for MGIT samples. DNA extraction was performed with mechanical ribolysis followed by purification using DiaSorin Liaison Ixt (DiaSorin, Italy) or DNeasy blood & tissue kit (Qiagen, Germany)(41). Target enrichment was performed using SureSelect with a custom-designed bait set providing coverage of the entire \textit{M. tuberculosis} genome as described previously(27). Batches of 48 multiplexed samples were sequenced on a NextSeq (Illumina, CA, USA).

**Bioinformatic analysis**

Bioinformatic analysis was performed with CLC Genomics Workbench v11.0 (Qiagen, Germany). DNA sequence reads were aligned to an H37Rv reference genome as detailed in the Supplementary Material Methods section (GenBank accession NC_000962.3). All samples had minimum 98% 1x reference genome coverage and mean coverage depth 40x across the genome. Variants falling within or near hypervariable elements were excluded (Supplementary Material: Table 1). A consensus sequence was extracted and used to determine the genotypic drug susceptibility profile. To construct the maximum likelihood phylogenetic tree, variants were called against the reference genome using VarScan v2.3.9 (Supplementary Material: Methods).

For the initial analysis of genetic diversity, variants were included if supported by \(\geq 2\) reads, with \(\geq 1\) forward and reverse read. The minimum supporting read threshold was increased in
a stepwise fashion from 2 to 20. Further analyses were performed on variant tracks where
variants were supported $\geq 5$ supporting reads including $\geq 1$ forward and reverse read.

To compare diversity between paired samples, we first mapped reads to the reference
genome using bwa mem v0.7.12[42]. After verifying all samples had adequate coverage with
qualimap[43] (mean $\pm$ standard deviation coverage at 10x: 98.0 $\pm$ 1.8%) and realigning
indels, variants were called with HaplotypeCaller in GATK v3.3.0[44] (Supplementary
Material: Methods). The gvcf files were combined for each pair of samples with
CombineGVCFs in GATK then screened to remove sites in variable regions and rRNA genes
with vcfintersect in vcflib, resulting in 39 paired gvcf files containing allele depths at variable
positions for diversity analysis.

Metagenomic assignment

Sequencing reads were classified using Kraken v0.10.6[45] against a custom Kraken
database previously constructed from all available RefSeq genomes for bacteria, archaea,
viruses, protozoa, and fungi, as well as all RefSeq plasmids (as of September 19th 2017) and
three human genome reference sequences[46]. The size of the final database after shrinking
was 193 Gb, covering 38,190 distinct NCBI taxonomic IDs.

To assess the proportion of contaminating reads that could generate spurious diversity
when mapped to M. tuberculosis ribosomal genes, we randomly subsampled 100 reads
taxonomically assigned as non-M. tuberculosis and performed a BLAST search with blastn
v2.2.28[47] against rRNA genes from the H37Rv reference genome. We only analysed hits of
at least 30 bases.
Statistics

Statistical analyses were performed with Prism v7.0 (Graphpad, CA, USA). The number of HAs in paired samples were compared using a two-tailed Wilcoxon matched-pairs signed rank test. Numbers of HAs found between groups were compared with chi-squared.

Within-sample diversity ($H$) was calculated using Shannon diversity from the allele frequencies ($p$). The Shannon index ($H_n$) expresses the positional entropy at each position ($n$), with the mean positional entropy ($\bar{H}$) indicating greater within-sample diversity:

$$H = \sum_n H_n = \sum_n \sum_{i \in \{A,C,G,T\}} p_{n,i} \log (p_{n,i})$$

To make a fair comparison for each pair of samples, after removing indels, sites were included if they contained a variant in at least one and had a depth coverage $\geq 30$ in both. We found that the depth coverage cutoff had no qualitative effect on the conclusions. The difference in mean within-sample diversity depending on DNA source was compared with a two-tailed Wilcoxon matched-pairs signed rank test.

Abbreviations

<table>
<thead>
<tr>
<th>DST</th>
<th>drug susceptibility testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>HA</td>
<td>heterozygous allele</td>
</tr>
<tr>
<td>MDR-TB</td>
<td>multidrug resistant-tuberculosis</td>
</tr>
<tr>
<td>MGIT</td>
<td>mycobacterial growth indicator tube</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>RAV</td>
<td>resistance-associated variant</td>
</tr>
<tr>
<td>rRNA</td>
<td>ribosomal RNA</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td>TB</td>
<td>tuberculosis</td>
</tr>
<tr>
<td>WGS</td>
<td>whole genome sequencing</td>
</tr>
</tbody>
</table>
Declarations

Ethics approval and consent to participate

All patients gave written informed consent to participate in the study. Ethical approval for the London study was granted by NHS National Research Ethics Service East Midlands–Nottingham 1 (reference 15/EM/0091). Ethical approval for the Durban study was granted by University of KwaZulu-Natal Biomedical Research Ethics Committee (reference BE022/13).

Consent for publication

Not applicable

Availability of data and materials

Original fastq files are available at NCBI Sequence Read Archive with BioProject reference PRJNA486713: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA486713/

Competing interests

The authors declare that they have no competing interests.

Funding

Camus Nimmo is funded by a Wellcome Trust Research Training Fellowship reference 203583/Z/16/Z. This work was additionally funded by National Institute for Health Research via the UCLH/UCL Biomedical Research Centre (grant number BRC/176/III/JB/101350) and the PATHSEEK European Union’s Seventh Programme for research and technological
development (grant number 304875). The funding bodies had no input on study design, analysis, data interpretation or manuscript writing.

Authors' contributions

Study conception: JB, ASP

Data collection: CB, KB

Analysis and interpretation: CN, LPS, RD, RW

Drafting of manuscript: CN, LPS

Revision of manuscript: FB, JB, ASP

Final approval of manuscript: CN, LPS, RD, RW, KB, CB, JB, FB, ASP

Acknowledgements

The authors would like to thank Sashen Moodley, Ashantha Govender and Colin Chetty in the Microbiology Core, Africa Health Research Institute.
### Table 1. Baseline patient characteristics for 39 patients (or as otherwise specified where data were missing). *Data missing for 1 patient.

<table>
<thead>
<tr>
<th></th>
<th>Count / mean</th>
<th>Range/percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td>36.7</td>
<td>22 – 64</td>
</tr>
<tr>
<td><strong>Male sex</strong></td>
<td>23/37</td>
<td>59.0%</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>3/35</td>
<td>8.6%</td>
</tr>
<tr>
<td>Black African</td>
<td>28/35</td>
<td>80.0%</td>
</tr>
<tr>
<td>Caucasian</td>
<td>4/35</td>
<td>11.4%</td>
</tr>
<tr>
<td><strong>HIV positive</strong></td>
<td>19/38</td>
<td>50.0%</td>
</tr>
<tr>
<td>CD4 count (median)</td>
<td>296.5*</td>
<td>17 – 707</td>
</tr>
<tr>
<td>On antiretroviral therapy at time of diagnosis</td>
<td>8/19**</td>
<td>42.1%</td>
</tr>
<tr>
<td>Drug</td>
<td>Resistance by phenotypic DST</td>
<td>Resistance by genotypic DST</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td><strong>First-line drugs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rifampicin</td>
<td>24/37 (64.9%)</td>
<td>24/39</td>
</tr>
<tr>
<td>Isoniazid</td>
<td>25/37 (67.6%)</td>
<td>24/39</td>
</tr>
<tr>
<td>Ethambutol</td>
<td>8/37 (21.6%)</td>
<td>17/39</td>
</tr>
<tr>
<td><strong>Second-line drugs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ofloxacin</td>
<td>0/24 (0.0%)</td>
<td>1/24</td>
</tr>
<tr>
<td>Kanamycin</td>
<td>1/24 (4.2%)</td>
<td>1/24</td>
</tr>
</tbody>
</table>

Table 2. Phenotypic and genotypic drug susceptibility testing (DST) results and sensitivity and specificity of genotypic DST relative to phenotypic DST. Phenotypic DST available for first-line drugs for 37 of the 39 patients, and for second-line drugs for 24 patients who demonstrated rifampicin drug resistance. *In two directly-sequenced sputum samples rifampicin RAVs were missed due to low coverage, although they were identified in the corresponding MGIT sample. **This sample had <1% of colonies grow in the presence of ofloxacin, so is categorised as sensitive but may have low-level or heteroresistance to fluoroquinolones (see main text).
<table>
<thead>
<tr>
<th>Gene</th>
<th>Heterozygous Allele Count</th>
<th>Gene length (base pairs)</th>
<th>Hypothesised gene function</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Shared</td>
<td>MGIT only</td>
<td>Sputum only</td>
</tr>
<tr>
<td></td>
<td>rrl</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Rv1319c</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>lppB</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Rv2561</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Rv3424c</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Rv2082</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>ppsA</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Rv1435c</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>ponA1</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Rv2277c</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>vapC31</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Rv2823c</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>lppA</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>pks12</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 3. Genes with the most heterozygous alleles (HAs) identified across the entire dataset.
### All variants vs H37Rv (fixed or heterozygous)

<table>
<thead>
<tr>
<th></th>
<th>Shared variants</th>
<th>MGIT only variants</th>
<th>Sputum only variants</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total variants</strong></td>
<td>33 153</td>
<td>1162</td>
<td>1217</td>
<td>35532</td>
</tr>
</tbody>
</table>

### Variants vs H37Rv present as heterozygous alleles (HAs) only

<table>
<thead>
<tr>
<th></th>
<th>MGIT 645 (1.9%)</th>
<th>Sputum 1074 (3.2%)</th>
<th>821 (67.5%)</th>
<th>2048 (5.8%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total variants</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Median HAs per sample</strong></td>
<td>21</td>
<td>3</td>
<td>15</td>
<td>40</td>
</tr>
</tbody>
</table>

### Variant type (% all HAs)

<table>
<thead>
<tr>
<th></th>
<th>SNP</th>
<th>MNP</th>
<th>Insertion</th>
<th>Deletion</th>
<th>Replacement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>500 (46.6%)</td>
<td>12 (1.1%)</td>
<td>303 (28.2%)</td>
<td>259 (24.1%)</td>
<td>0 (0.0%)</td>
</tr>
</tbody>
</table>

### Coding change (% all HAs)

<table>
<thead>
<tr>
<th></th>
<th>Non-synonymous</th>
<th>Synonymous</th>
<th>Intergenic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>395 (36.8%)</td>
<td>159 (14.8%)</td>
<td>520 (48.4%)</td>
</tr>
</tbody>
</table>

### Non-synon/synon ratio

|                      | 2.48 | 2.47 | 1.86 | 2.19 |

### Stop codon (% of all non-synonymous HAs)

|                      | 4 (1.0%) | 1 (1.3%) | 9 (2.8%) | 14 (1.8%) |

### Frameshift (% of all non-synonymous HAs)

|                      | 185 (46.8%) | 19 (24.1%) | 47 (14.8%) | 251 (31.7%) |
Table 4. Variants identified in MGIT derived, sputum derived, or both sequences from paired samples. Values given represent totals for the 39 paired samples. SNP = single nucleotide polymorphism; MNP = multi-nucleotide polymorphism.
<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Phenotypic resistance</th>
<th>Mutation</th>
<th>Frequency (MGIT/sputum)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1002</td>
<td>Rifampicin</td>
<td>rpoB S450L</td>
<td>100%/100%</td>
<td>High confidence resistance mutation</td>
</tr>
<tr>
<td>F1002</td>
<td>Rifampicin</td>
<td>rpoC G332R(48)</td>
<td>82.6%/21.7%</td>
<td>Putative compensatory mutations</td>
</tr>
<tr>
<td>F1002</td>
<td>Rifampicin</td>
<td>rpoC L516P(48)</td>
<td>12.7%/7.7%</td>
<td>Putative compensatory mutations</td>
</tr>
<tr>
<td>F1002</td>
<td>Rifampicin</td>
<td>rpoC P1040S(49)</td>
<td>21.7%/12.3%</td>
<td>Putative compensatory mutations</td>
</tr>
<tr>
<td>F1007</td>
<td>Isoniazid (high)</td>
<td>ahpC c-52t(38)</td>
<td>60.0%/50.7%</td>
<td>Rare, have been associated with resistance</td>
</tr>
<tr>
<td>F1007</td>
<td>Isoniazid (high)</td>
<td>ahpC g-48a(38)</td>
<td>28.6%/30.3%</td>
<td>Rare, have been associated with resistance</td>
</tr>
<tr>
<td>F1061</td>
<td>Rifampicin</td>
<td>rpoB H445D</td>
<td>16.1%/0.0%*</td>
<td>High confidence resistance mutation</td>
</tr>
<tr>
<td>F1061</td>
<td>Rifampicin</td>
<td>rpoB S450W</td>
<td>84.4%/0.0%*</td>
<td>High confidence resistance mutation</td>
</tr>
</tbody>
</table>
| F1066      | Isoniazid (high)     | katG N218fs | 0.0%/6.9%            | \n
Possible resistance mutations, not previously described

RF021      | Pyrazinamide – testing failed | pncA Q122H | 0%/2.5%              | Putative compensatory mutations, not previously described |
Table 5. Resistance-associated variants present as heterozygous alleles (HAs). *These mutations were also present in sputum but due to low coverage of the area (3 and 4 reads respectively) variant calling criteria were not met.
Figure legends

Figure 1. Variation in total number of heterozygous alleles (HAs) identified across all 39 patients in sequences generated from sputum and MGIT depending on minimum supporting read count threshold. Direct sputum samples indicated by red squares, MGIT samples blue circles.

Figure 2. (A) Number of heterozygous alleles (HAs) found in directly sequenced sputum only (sputum), MGIT (MGIT) only or in both samples (shared) by patient. (B) Mean Shannon diversity at variable positions across pairs of samples (∙) as calculated for MGIT and sputum-derived sequences. Size of point indicates number of variable positions considered (see Methods).
References


Minimum read count vs. Heterozygous alleles for MGIT and Sputum samples.

- **MGIT**
- **Sputum**

The graph shows the decrease in heterozygous alleles as the minimum read count increases for both MGIT and Sputum samples. The y-axis represents the number of heterozygous alleles, while the x-axis represents the minimum read count.
A

![Graph showing heterozygous alleles across patient IDs for Sputum and MGIT.](image)

B

![Scatter plot showing mean Shannon diversity of variant positions (Sputum) vs. MGIT.](image)

Number of variants across paired samples:
- 10,000
- 20,000
- 30,000

---

Additional text: (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.