Rapid Whole Genome Sequencing of *M. tuberculosis* directly from clinical samples

Amanda C. Brown[1,2] *, Josephine M. Bryant[3] *, Katja Einer-Jensen[4], Jolyon

Holdstock[1], Darren T Houniet[1], Jacqueline Z.M. Chan[1], Daniel P. Depledge[3],

Vladyslav Nikolayevskyy[5], Agnieszka Broda[5], Madeline J. Stone[6], Mette T.

Christiansen[3], Rachel Williams[3], Michael B. McAndrew[1], Helena Tutill[3], Julianne

Brown[3], Mark Melzer[7], Caryn Rosmarin[7], Timothy D. McHugh[8], Robert J. Shorten[8,9],

Francis Drobniewski[5], Graham Speight[1], Judith Breuer[3]

Study Group: PATHSEEK consortium

1. Oxford Gene Technology, Begbroke Science Park, Oxford OX5 1PF
2. Current address: Department of Microbiology and Immunology, Cornell University, Ithaca, NY 14853 USA
3. UCL, Division of Infection and Immunity, The Cruciform Building, 60 Gower St, London WC1E 6BT
4. Qiagen-AAR, Silkebrogvej 3, Aarhus, Denmark
5. National Mycobacterium Reference Laboratory (NMRL), Abernethy House, ICMS, 2 Newark St, London E1 2AT
6. Dept. Microbiology, Frimley Health NHS Foundation Trust, Wexham Park Hospital, Berkshire, SL2 4HL
7. Barts Health NHS Trust, West Smithfield, London, EC1A 7BE
8. Centre for Clinical Microbiology, UCL, Royal Free Campus, Rowland Hill St, London, NW3 1QG
9. Specialist Microbiology Network, Public Health Laboratory Manchester, Manchester Royal Infirmary, Oxford Road, Manchester M13 9WL

* Authors contributed equally to this work

Author for correspondence: Josephine M. Bryant

j.bryant@ucl.ac.uk

UCL, Division of Infection and Immunity, The Cruciform Building, 60 Gower St,

London WC1E 6BT

Tel: +44(0)7525924037

Abstract

The rapid identification of antimicrobial resistance is essential for effective treatment of highly resistant *Mycobacterium tuberculosis* (*M. tb*). Whole genome sequencing provides comprehensive data on resistance mutations and strain typing for monitoring transmission, but unlike conventional molecular tests, this has only previously been achievable from cultured *M. tb*. Here we describe a method utilising biotinylated RNA baits, designed specifically for *M. tb* DNA to capture full *M. tb* genomes directly from infected sputum samples, allowing whole genome sequencing without the requirement of culture. This was carried out on 24 smear-positive sputum samples, collected from the UK and Lithuania where a matched culture sample was available, and two samplesthat had failed to grow in culture. *M. tb* sequencing data was obtained directly from all 24 smear-positive culture-positive sputa, of which 20 were high quality (>20x depth and >90% of genome covered). Results were compared with conventional molecular and culture-based methods, and high levels of concordance were observed between phenotypical resistance and predicted resistance based on genotype. High quality sequence data was obtained from one smear positive culture negative case. This study demonstrates for the first time, the successful and accurate sequencing of *M. tb* genomes directly from uncultured sputa. Identification of known resistance mutations within a week of sample receipt offers the prospect for personalised, rather than empirical, treatment of drug resistant tuberculosis, including the use of antimicrobial-sparing regimens, leading to improved outcomes.

Funding

Introduction

The global incidence of multi, extensively and totally drug resistant tuberculosis has risen over the last decade (1), making it increasingly important to rapidly and accurately detect resistance. The gold standard for antimicrobial resistance testing relies on bacterial culture, which for *M. tb* can take upwards of several weeks. Molecular tests, such as the Xpert (MTB/RIF) and line probe assays, which can be used directly on sputum have improved identification of multi-drug resistant (MDR) *M.tb* but are only able to identify limited numbers of specific resistance mutations (2, 3).

Whole bacterial genome sequencing (WGS) allows simultaneous identification of all known resistance mutations as well as markers with which transmission can be monitored (4). WGS of *M.tb* provides superior resolution over other current methods such as spoligotyping and MIRU-VNTR for strain genotyping (5) and its usefulness in defining outbreaks has been demonstrated (6-9). Currently however, WGS of *M. tb* requires prior bacterial enrichment by culturing and therefore most outbreak studies have been retrospective (6-8). Recently WGS of *M. tb* has been achieved from three-day old MGIT (Mycobacterial growth indicator tube) culture, thus reducing the time from sample receipt to resistance testing to less than a week (10). However,

79   with the mean time to positive MGIT culture being 14 days (11, 12), most WGS

80   results will not be available for more than two weeks, which is too long a delay

81   before starting therapy. Moreover, the extent to which even limited culture perturbs

82   the original sample composition remains unknown, especially in cases where a

83   patient is suffering from infection with multiple strains, a common occurrence in

84   developing countries where it has been observed in up to 19% of cases (13). As

85   described here we utilised the oligonucleotide enrichment technology SureSelectXT

86   (Agilent) method to obtain the first *M. tb* genome sequences directly from both smear

87   positive and smear negative sputum.

88

Methods

89

Samples

90

91   A total of 58 routine diagnostic samples from the UK and Lithuania, including 24

92   smear-positive sputum specimens from pulmonary TB patients and 24 matching

93   cultures (grown on Middlebrook 7H11 plates from the relevant sputum specimens,

94   see below), and 10 sputum samples from patients who had previously been

95   diagnosed with TB and which failed to grow in culture, were analysed. Further details

96   can be found in supplementary table 1. Sputum was visually scored as 1+ to 3+ for

97   acid fast bacilli (AFBs). Sequencing and subsequent analysis were processed blind

98   with respect to smear and resistance results.

Bacteriological methods

99

100   Prior to treatment, all sputum specimens were kept frozen at -20 ˚C. Bacteriological

101   culture samples were processed as follows. Samples were decontaminated using N-

4

acetyl-L-cysteine/NAOH (1% NaOH final concentration) and re-suspended after centrifugation in 2 mL phosphate buffer (pH 6.8). Subsequently, 0.1 ml of the suspension was used for inoculation onto Middlebrook 7H11 media while the remaining suspension was used for the genomic DNA extraction directly from sputum (see below). Plates were incubated at 37 °C for at least four weeks or until visible growth was obtained.

DNA extraction from sputum

The bacterial suspension used for inoculation was re-pelleted by centrifugation at 16,000 g. Supernatants were decanted, and pelleted cells re-suspended in 0.3 mL Tris-EDTA (TE), buffer and transferred to sterile 2 mL screw caps tubes containing ~250 µL 0.1 mm glass beads (Becton Dickinson). Microorganisms were heat killed at 80°C for 50 minutes and then frozen at -20°C; after thawing the tubes were vortexed for three minutes and centrifuged for five minutes at 16,000 g. The supernatant was transferred to a clean 2mL tube for subsequent DNA purification using the DNeasy Blood and Tissue DNA extraction Kit (Qiagen) as per manufacturer's instructions. Genome copies were measured in the sputum samples using the Artus® *M. tuberculosis* RG PCR Kit (Qiagen), as per manufacturer's instructions.

DNA extraction from cultures

Two loopfuls of *M.tb* growth from Middlebrook 7H11 plates were transferred into 2 mL screw caps tubes containing ~250 µL of 0.1mm glass beads (Becton Dickinson) and 0.3 mL TE buffer. Subsequent processing, genomic DNA extraction and purification were done as described for the sputum samples.

Resistance profiling

125 All isolates were tested for susceptibility to first line drugs rifampicin (RIF), isoniazid

126 (INH), ethambutol (EMB), pyrazinamide (PZA), and streptomycin (STR). Isolates

127 resistant to at least RIF and INH (i.e. multidrug resistant, MDR-TB) were additionally

128 tested for susceptibility to kanamycin (KAN), amikacin (Amk), ofloxacin (OFL),

129 capreomycin (CAP), ethionamide (ETH), prothionamide (PTH), and par-

130 aminoasalicylate sodium (PAS).

131

132 Drug susceptibility testing (DST) was carried out on an automated liquid media-

133 based system Bactec MGIT960 (Becton Dickinson) using standard drug

134 concentrations (in micrograms per millilitre) as follows: STR 1.0; INH 0.1; RIF 1.0;

135 EMB 5.0; PZA 100.0; OFL 2.0; Amk 1.0; CAP 2.5; KAN 5.0; ETH 5.0; PTH 2.5; and

136 PAS 4.0.(14)

137 <u>Spoligotyping</u>

138 Spoligotyping was carried out as described previously using membranes with

139 immobilised oligonucleotide probes (Ocimum Biosolutions) (15). For identification of

140 genetic families and lineages, 43-digit binary spoligotyping codes were entered into

141 MIRU-VNTRplus database (www.miru-vntrplus.org) and families identified using

142 similarity search algorithm.

143 <u>SureSelect^{XT} Target Enrichment: RNA baits design</u>

144 120-mer RNA baits spanning the length of the positive strand of H37Rv *M. tb*

145 reference genome (AL123456.3),(16) were designed using an in-house Perl script

146 developed by the PATHSEEK consortium. The specificity of the baits was verified by

147 BLASTn searches against the Human Genomic + Transcript database. The custom

148 designed *M. tuberculosis* bait library was uploaded to SureDesign and synthesised

149 by Agilent Technologies.

150

151 SureSelect^{XT} Target Enrichment: Library preparation, hybridisation and Illumina

152 sequencing.

153 Prior to processing *M.tb* DNA samples were quantified and carrier human genomic

154 DNA (Promega) was added to obtain a total of 3 µg DNA input for library preparation.

155 All DNA samples were sheared for 4x60 seconds using a Covaris S2 (duty cycle

156 10%, intensity 4 and 200 cycles per burst using frequency sweeping). The samples

157 were then subjected to library preparation using the SureSelect^{XT} Target Enrichment

158 System for Illumina Paired-End Sequencing Library protocol (V1.4.1 Sept 2012).

159 Prior to hybridisation eight cycles of pre-capture PCR was used, and ~750 ng of

160 amplified product was included in each hybridisation (24 hours, 65 ˚C). 16 cycles of

161 post capture PCR was performed, with indexing primers. The resulting library was

162 run on a MiSeq (Illumina) using a 600bp reagent kit, typically in pools of 8 or 10,

163 some sputum smear positive 1+ samples were run in smaller pools to increase

164 coverage. Base calling and sample demultiplexing were generated as standard on

165 the MiSeq machine producing paired FASTQ files for each sample. The raw

166 sequencing data has been deposited on the European Nucleotide archive (upon

167 acceptance of publication). An overview of the process is presented in

168 Supplementary figure 1.

169 Sequence Analysis

170 The samples were analysed using a reference based mapping approach

171 implemented in CLC Genomics workbench (v. 7.5). Prior to mapping the reads were

172 trimmed to remove low quality sequence at the end of reads or adaptor

173 contamination. The reads were mapped against the H37Rv genome (AL123456.3)

174 using default parameters with the addition of a similarity threshold to remove non-

175 *M.tb* reads, by which any reads where at least 90% of the length does not match the

176 reference by at least 90% were discarded. This was required to remove non-*M.tb*

177 reads. Duplicate reads were then removed from the mapped reads, and the average

178 depth of coverage calculated. The percentage of on-target reads (OTR) was

179 calculated by counting the number of reads that were successfully mapped to

180 H37Rv. Any reads that did not map were assumed to be off-target (not *M.tb*).

181 Bases were called using VarScan (v 2.3.7),(17) applying high stringency parameters

182 including a minimum depth of 4 reads, a minimum average quality of 20, a p value

183 cutoff of 99e-02 and an absence of heterozygosity at a level greater than 10%. A

184 consensus sequence was generated where only called bases were considered, and

185 any bases which failed quality thresholds were called as Ns. To build the phylogeny

186 any variants which were identified in IS elements or the PE, PPE gene families were

187 excluded, as these regions are recognized to be prone to false positive SNP calls.(8)

188 The remaining positions (representing 92% of the genome) were then used to build a

189 maximum likelihood tree using RAxML v 8.0.0(18) with 100 bootstrap replicates.

190 Genome coverage was calculated by dividing the number of high quality bases

191 successfully called (as per VarScan above) by the reference genome (H37Rv) size.

192 Depth of coverage refers to the number of reads supporting a position.

193 <u>Calling genotypic resistance</u>

194 Potential drug susceptibility associated variants were detected using a custom Perl

195 script using positions identified in a curated drug resistance database

196　(http://pathogenseq.lshtm.ac.uk/rapiddrdata) (19) from bam and bcf files (20).

197　Variants were considered if they were supported by at least 2 forward and reverse

198　reads, had p values of at least 0.05 for strand bias, and 0.001 for read end bias,

199　base quality bias and mapping quality bias as calculated by bcftools (20). A sample

200　was called as genotypically resistant if it had a mutation in over 10% of reads. Any

201　mutation identified in the ribosomal RNA genes were inspected manually to exclude

202　any that may be the result of off-target enrichment of these highly conserved regions.

203　Those that were found on reads that formed distinct haplotypes, where variants were

204　found in close association with other variants on multiple reads were excluded as

205　they likely belonged to non-*M. tb* species.

206　The analysis was also carried out independently on a customized version of the CLC

207　Genomics Workbench (QIAGEN-AAR), which facilitates a fully-automated pipeline

208　including the steps of trimming, mapping to reference, removal of duplicate mapped

209　reads, variant calling and cross-referencing with the resistance database (described

210　above). Variants called using the automated workflow (using the Low Frequency

211　Variant Detector, CLC Genomic Workbench), were considered significant if the

212　average quality was above 30, a frequency greater than 10% and the forward and

213　reverse read balance was above 0.35. Variants were inspected manually for possible

214　contamination. The runtime when using a standard laptop (Macbook Pro) was on

215　average 1h per sample. The resistance genotypes called were in agreement with

216　those identified using the workflow described above

217

218　**Results**

219

<u>Successful enrichment directly from sputum in both smear positive and negative</u>

<u>tuberculosis cases</u>

To assess the potential benefits of enrichment strategies for WGS of clinical *M. tb*, we compared the percentage of on-target reads (%OTR), as defined in the methods, and the mean sequencing depth) for two sputum samples, each processed with and without enrichment. The average %OTR for *M. tb* sequenced directly (no-enrichment) from sputum was 0.3%, with 4.6x sequencing depth, compared to with enrichment which generated a %OTR of 82%, with a mean depth of 200x (Figure 1). Although cultured *M. tb* sequenced well with and without enrichment, the former gave greater mean read depth (Supplementary Figure 2). Even coverage across the genome was obtained with no bias observed for particular regions or genes (Supplementary figure 3).

Over 98% of the *M. tb* genome was recovered from 20 of 24 (83%) smear-positive culture-positive sputa. Fully complete genomes are not achievable for *M. tb* using short-read sequencing technology, due to the difficulties presented by repetitive regions and the PE and PPE genes (8, 21). Similar levels of genome coverage and sequencing depth were also obtained from the non-enriched matched cultures (Figure 2a). The depth of coverage obtained for four sputum samples was poor (MTB-41: 9x, MTB-42: 14x, MTB-43: 6x and MTB-44: 11x) resulting in a genome coverage of less than 90%. In the case of MTB-43 and MTB-44, which had an input of 1 and <1 genome copies per µl respectively, (these values are out of range of reliable standards as measured by real-time PCR) this is likely to be due to a low pathogen load.

In addition, we were able to enrich and sequence *M.tb* from two smear-positive but culture negative sputum samples (MTB-69 and MTB-73), successfully recovering sequence data from the former with a sequence depth of 8.5x (Figure 2B). We also attempted to sequence eight culture-negative smear-negative sputum samples, which were obtained from previously diagnosed patients (MTB-67-72 and MTB-74 - 76). Surprisingly, for two of these samples, MTB-68 and MTB-76, we obtained high quality *M. tb* sequence data with an average depth of coverage of 9x and 22x, respectively (Figure 2b). For five of the samples we detected low numbers of *M.tb* reads (<1x depth of coverage), which may represent a very low load or residual DNA from dead bacilli. No full length *M. tb* reads were detected in the final sample (MTB-72).

Concordance between genotypes obtained from culture and sputa matched pairs

Using the high quality variable sites called we constructed a maximum likelihood phylogenetic tree. Six samples, with less than 90% genome and SNP position coverage, had an unusual phylogenetic positioning, close to nodes on the tree (Supplementary figure 4), a pattern consistent with a lack of informative sites. With these samples excluded, the resulting robust phylogeny revealed that for all of the matched pairs an identical or near-identical genome was obtained from culture and sputa (Supplementary figure 5).

Concordance between resistance phenotype and genotype

For the 24 matched pairs we sought to identify the genetic resistance determinants which could explain their antibiotic susceptibility profile. Predicted resistance mutations were 100% concordant between sequences obtained from culture and sputa (Table 1), with the exception of the low coverage samples (MTB-41, MTB-42,

267   MTB-43 and MTB-44) for which we were unable to confidently call variants at many
268   of the targeted loci.

269   The predicted resistance genotype agreed well with the phenotypic resistance
270   profiles, with a possible resistance conferring mutation being detected in 88%
271   (59/67) of phenotypically resistant cases, and no known resistance mutation in 94%
272   (72/77) of sensitive cases. Two phenotypically pyrazinamide resistant cases
273   (isolated from Patients 7 and 14) both belonging to the URAL lineage were identified
274   as having a large chromosomal deletion (8.64kb) resulting in the removal of the *pncA*
275   gene, the activator of the pro-drug pyrazinamide, plus ten surrounding genes
276   (Supplementary figure 6).

277   Four samples were phenotypically ethambutol sensitive but had a mutation in codon
278   306 of the *embB* gene. This mutation has been observed to cause both low and
279   high level resistance to ethambutol, so may lead to a borderline phenotype (22, 23).
280   Patient 15's isolate was phenotypically ethambutol sensitive but had a Q497R
281   mutation in the *embB* gene, which has previously been associated with being
282   sensitive in both clinical isolates (24) and through the construction of isogenic
283   mutants (25) so was discounted. Similarly patient 2's isolate was phenotypically
284   sensitive to isoniazid, but had a G269S mutation in the *kasA* gene, which has also
285   previously been found in sensitive isolates (26). Patient 5's isolate was also
286   phenotypically sensitive to rifampicin notwithstanding a L452P (codon 533 in *E. coli*)
287   mutation in the *rpoB* gene. This mutation has also been associated with both high
288   and low rifampicin resistance in the literature (27). *M. tb* isolated from this patient in
289   the past had been found to be rifampicin resistant suggesting that either this
290   mutation results in a borderline phenotype. Alternatively, a mixture of rifampicin
291   resistant and sensitive strains could have been present in this patient, although this

292   was not detected in the sequencing data obtained from either the sputum or culture.

293   The remaining eight samples were phenotypically resistant, with an absence of any

294   described or speculative causative genetic mutations. Five of these were

295   phenotypically resistant to second-line drugs for which the genetic basis of

296   resistance is less well understood. These discrepancies highlight that the current

297   limitation on our ability to detect resistance via whole genome sequencing is not the

298   detection itself, but rather lack of data on the genetic correlates of resistance.

299   Any alleles detected at a low level (<10%), were excluded from this analysis due to

300   the potential problem of carry-over on the sequencing platform which has been

301   previously described (28). Further work will be required to quantify the validity of

302   these mutations, or to assess their clinical significance. In the majority of cases,

303   resistance alleles had reached fixation or near-fixation in both culture and sputa

304   samples, as they were found in 98-100% of the reads. In patient 10 however,

305   significant heterozygosity was detected, with more than one allele being detected at

306   greater than 10% at a single position. A mixture of three different resistance alleles

307   and one sensitive allele were detected within a single codon of the *gyrA* gene (Figure

308   3). Remarkably almost identical proportions were detected in the corresponding

309   culture sample.

310

311

312   **Discussion**

313   Whole genome sequencing of bacteria has been shown to provide comprehensive

314   data on antimicrobial resistance, which could be used to inform antimicrobial

315 prescribing. However current methods which rely on culturing the organism prior to

316 sequencing are slow and so of limited use in patient management. As a result initial

317 antimicrobial prescribing for resistant *M. tb*, remains largely empirical in the early

318 phase of treatment. Currently MDR *M. tb* can be diagnosed rapidly on the Xpert

319 MTB/RIF system, but a rapid test for extensively drug resistant (XDR) cases is

320 unavailable. As Xpert (MTB/RIF) focuses only on *rpoB* (RIF) mutations unusual

321 resistance patterns where strains are rifampicin sensitive but show other resistance,

322 such as the isoniazid resistant, rifampicin sensitive case included in this study, are

323 missed. Here, we describe the recovery and sequencing of near-complete genomes

324 directly from 81% (21/26) of smear positive sputa, including those staining for low

325 numbers (+1) of Acid Fast Bacilli (AFBs), within a timescale (up to 96 hours) that

326 could allow personalised antimicrobial treatment for both sensitive and resistant

327 cases, including XDR TB.

328 *M. tb* is particularly appropriate for the use of diagnostic WGS with enrichment, as,

329 unlike the majority of pathogenic organisms, *M. tb* has a well characterised clonal

330 nature, with relatively low levels of sequence variation and does not undergo

331 recombination or horizontal transfer (29), thus a stable set of oligonucleotide baits

332 can be created and sequence data can be mapped against a reference genome. We

333 have demonstrated that enrichment of *M. tb* provides sequencing data that matches

334 the quality and quantity of data obtained via sequencing from culture. Moreover, we

335 were able to recover high quality *M. tb* sequencing data from one smear-positive and

336 one smear-negative case, both from cases who had received anti-TB therapy and

337 which both failed to grow in culture. However without further clinical information it is

338 difficult to interpret these cases. Smear-positive culture-negative cases are most

339 commonly thought to be due to the on-going persistence of dead bacilli in sputum

340 samples (30). For this reason, previously treated cases are currently not
341 recommended for use on PCR-based diagnostic systems such as Xpert, that cannot
342 distinguish between dead or live bacilli. Further investigation will be required to
343 assess the suitability of targeted enrichment in the context of different clinical
344 scenarios. There were four smear-positive culture-positive cases where less optimal
345 data were obtained from sputa, although we envisage that sequencing of such low
346 titre samples could be improved through further optimization or increased
347 sequencing depth. It is worth noting these samples were deemed failures based on
348 commonly used SNP calling thresholds employed by others in the field. Further work
349 will be required to robustly establish parameters that are sufficient for clinical use
350 and interpretation, particularly when considering low frequency variants.

351 Sequencing directly from the clinical sample may reduce any possible biases
352 associated with culture. The overall presence of hetero-resistance in this study was
353 low (one patient), with most resistance conferring mutations observed as close to
354 fixation, i.e. the entire sampled population is resistant. However, in endemic settings
355 mixed infections have been observed to be much more prevalent especially in HIV
356 positive patients (14, 31-33). The detection of these hetero-resistant cases is not
357 only important for our understanding of how resistance evolves, but could impact on
358 clinical management (34). Further studies are required to explore any bias on
359 genetic diversity that may be introduced by culture, particularly in the context of
360 mixed-strain infections.

361 A disadvantage of the approach presented here is that it is relatively expensive:
362 currently costing approximately $350 (USD) per sample in our laboratory. It also
363 requires skills and machinery currently not available in most microbiological
364 laboratories. An alternative and cheaper rapid sequence based approach would be

365 to deep sequence total DNA from sputa samples without enrichment.. A recent study

366 found they could recover *M. tb* reads from eight smear and culture positive samples

367 (35). However, in agreement with our study, they obtained a very low depth of

368 coverage (<1x) in the absence of enrichment, so the usefulness of this approach is

369 likely to be limited to detection, and is unlikely to provide the detailed genotype and

370 resistance information that is presented here in a high-throughput manner.

371 In summary, we have demonstrated whole *M. tb* genome sequencing directly from

372 smear positive, culture positive sputa within a clinically relevant time frame that

373 would enable pro-active patient management. The quality of sequence data allowed

374 us to accurately call mutations that are known to be associated with resistance to

375 first and second line drugs. Furthermore, excluding the need for culture affords new

376 opportunities for biological insights into the evolution of *M. tb* antimicrobial resistance

377 and within-patient evolution.

378

379 **Figures**

380 Figure 1: Mean coverage and percentage of on target reads (OTR) when sequencing

381 from sputum with and without enrichment for two samples

382 Figure 2: (A) Depth of coverage obtained for smear positive samples from sputum

383 and culture. (B) Depth of coverage for sputum sequence from smear positive

384 samples which failed to grow. Level of smear positivity is shown, with the remaining

385 being smear negative.

386 Figure 3: Heteroresistance in gyrA in patient 10. R= resistant allele with suffix

387 indicating codon position, S= absence of resistant allele.

388 Table 1: Resistance phenotype and genotype of matched pairs. R= a mutation exists

389 at greater than 10%. Low R= mutation in codon 306 of embB gene which is thought

390 to confer low level resistance to ethambutol. Rif = rifampicin, Inh= isoniazid, Emb =

391 ethambutol, Pza = pyrazinamide, Str = streptomycin, Ofl= Ofloxacin

392 (fluoroquinolones), Pas = para-Aminosalicylic acid, Amg = aminoglycosides, Thi =

393 thionamides.

394

395 <u>Contributions</u>

396 The manuscript was written by JMB, ACB and JBreuer with input from all other

397 authors. Bioinformatic analysis and pipeline development was carried out by JMB,

398 JH, DTH, ACB, JZMC, and KEJ. Samples were supplied and processed by FD CR,

399 MM, TDM, VN, AB, RJS, MS HT and JBrown. Enrichment and sequencing was

400 carried out by ACB, and JZMC. ACB, DPD and MTC and members of the

401 PATHSEEK consortium contributed to protocol optimisation. The study was co-

402 ordinated by RW. The study was conceived and initiated by the PATHSEEK

403 consortium and managed by J Breuer, MBM, and GS.

404

405 <u>Acknowledgments</u>

406 The authors would like to acknowledge Edita Pimkina (Vilnius University Hospital

407 Santariskiu Klinikos) for the collection and processing of samples from Lithuania. We

408 are grateful for Poul Liboriussen and Jens Johansen (QIAGEN-AAR) contributions to

409 development of the customised and automated pipeline based on CLC Genomic

410 Workbench. J Breuer is supported by the UCL/UCLH and J Brown by the

415   <u>Competing interests</u>

References

423

424   1.   **Gandhi NR, Nunn P, Dheda K, Schaaf HS, Zignol M, van Soolingen D, Jensen P,**
425   **Bayona J.** 2010. Multidrug-resistant and extensively drug-resistant tuberculosis: a
426   threat to global control of tuberculosis. Lancet **375:**1830-1843.
427   2.   **Weyer K, Mirzayev F, Migliori GB, Van Gemert W, D'Ambrosio L, Zignol M,**
428   **Floyd K, Centis R, Cirillo DM, Tortoli E, Gilpin C, de Dieu Iragena J, Falzon D,**
429   **Raviglione M.** 2013. Rapid molecular TB diagnosis: evidence, policy making and
430   global implementation of Xpert MTB/RIF. Eur Respir J **42:**252-271.
431   3.   **Drobniewski F, Nikolayevskyy V, Maxeiner H, Balabanova Y, Casali N,**
432   **Kontsevaya I, Ignatyeva O.** 2013. Rapid diagnostics of tuberculosis and drug
433   resistance in the industrialized world: clinical and public health benefits and barriers
434   to implementation. BMC Med **11:**190.
435   4.   **Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, Kontsevaya**
436   **I, Corander J, Bryant J, Parkhill J, Nejentsev S, Horstmann RD, Brown T,**
437   **Drobniewski F.** 2014. Evolution and transmission of drug-resistant tuberculosis in a
438   Russian population. Nature Genetics **46:**279-286.
439   5.   **Comas I, Homolka S, Niemann S, Gagneux S.** 2009. Genotyping of genetically
440   monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights
441   the limitations of current methodologies. PLoS ONE **4:**e7815.

442   6.   **Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S,**
443   **Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K,**
444   **Jones SJ, Brinkman FS, Brunham RC, Tang P.** 2011. Whole-genome sequencing
445   and social-network analysis of a tuberculosis outbreak. The New England journal of
446   medicine **364:**730-739.
447   7.   **Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW,**
448   **Wilson DJ, Hawkey PM, Crook DW, Parkhill J, Harris D, Walker AS, Bowden R,**
449   **Monk P, Smith EG, Peto TE.** 2013. Whole-genome sequencing to delineate
450   Mycobacterium tuberculosis outbreaks: a retrospective observational study. The
451   Lancet infectious diseases **13:**137-146.
452   8.   **Roetzer A, Diel R, Kohl TA, Ruckert C, Nubel U, Blom J, Wirth T, Jaenicke S,**
453   **Schuback S, Rusch-Gerdes S, Supply P, Kalinowski J, Niemann S.** 2013. Whole
454   genome sequencing versus traditional genotyping for investigation of a
455   Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study.
456   PLoS medicine **10:**e1001387.
457   9.   **Smit PW, Vasankari T, Aaltonen H, Haanperä M, Casali N, Marttila H, Marttila J,**
458   **Ojanen P, Ruohola A, Ruutu P, Drobniewski F, Lyytikäinen O, Soini H.** 2015.
459   Enhanced tuberculosis outbreak investigation using whole genome sequencing and
460   IGRA. Eur Respir J **45:**276-279.
461   10.   **Koser CU, Bryant JM, Becq J, Torok ME, Ellington MJ, Marti-Renom MA,**
462   **Carmichael AJ, Parkhill J, Smith GP, Peacock SJ.** 2013. Whole-genome
463   sequencing for rapid susceptibility testing of M. tuberculosis. The New England
464   journal of medicine **369:**290-292.
465   11.   **Fadzilah MN, Ng KP, Ngeow YF.** 2009. The manual MGIT system for the detection
466   of M. tuberculosis in respiratory specimens: an experience in the University Malaya
467   Medical Centre. Malays J Pathol **31:**93-97.
468   12.   **Chihota VN, Grant AD, Fielding K, Ndibongo B, van Zyl A, Muirhead D,**
469   **Churchyard GJ.** 2010. Liquid vs. solid culture for tuberculosis: performance and cost
470   in a resource-constrained setting. Int J Tuberc Lung Dis **14:**1024-1031.
471   13.   **Warren RM, Victor TC, Streicher EM, Richardson M, Beyers N, Gey van Pittius**
472   **NC, van Helden PD.** 2004. Patients with active tuberculosis often have different
473   strains in the same sputum specimen. American Journal of Respiratory and Critical
474   Care Medicine **169:**610-614.
475   14.   **Krüüner A, Yates MD, Drobniewski FA.** 2006. Evaluation of MGIT 960-based
476   antimicrobial testing and determination of critical concentrations of first- and second-
477   line antimicrobial drugs with drug-resistant clinical strains of Mycobacterium
478   tuberculosis. J Clin Microbiol **44:**811-818.
479   15.   **Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S,**
480   **Bunschoten A, Molhuizen H, Shaw R, Goyal M, van Embden J.** 1997.
481   Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for
482   diagnosis and epidemiology. J Clin Microbiol **35:**907-914.
483   16.   **Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV,**
484   **Eiglmeier K, Gas S, Barry CE, 3rd, Tekaia F, Badcock K, Basham D, Brown D,**
485   **Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N,**
486   **Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver**
487   **K, Osborne J, Quail MA, Rajandream MA, Rogers J, Rutter S, Seeger K, Skelton**
488   **J, Squares R, Squares S, Sulston JE, Taylor K, Whitehead S, Barrell BG.** 1998.
489   Deciphering the biology of Mycobacterium tuberculosis from the complete genome
490   sequence. Nature **393:**537-544.
491   17.   **Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA,**
492   **Mardis ER, Ding L, Wilson RK.** 2012. VarScan 2: somatic mutation and copy
493   number alteration discovery in cancer by exome sequencing. Genome Res **22:**568-
494   576.
495   18.   **Stamatakis A.** 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic
496   analyses with thousands of taxa and mixed models. Bioinformatics **22:**2688-2690.

497 19. **F C.** From genome to bedside: closing the gap with ultra-rapid analysis for anti-
498   tuberculosis drug resistance. . Under Review.
499 20. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis**
500   **G, Durbin R.** 2009. The Sequence Alignment/Map format and SAMtools.
501   Bioinformatics **25:**2078-2079.
502 21. **Bryant JM, Schurch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V,**
503   **Kremer K, van Hijum SA, Siezen RJ, Borgdorff M, Bentley SD, Parkhill J, van**
504   **Soolingen D.** 2013. Inferring patient to patient transmission of Mycobacterium
505   tuberculosis from whole genome sequencing data. BMC infectious diseases **13:**110.
506 22. **Plinke C, Walter K, Aly S, Ehlers S, Niemann S.** 2011. Mycobacterium tuberculosis
507   embB codon 306 mutations confer moderately increased resistance to ethambutol in
508   vitro and in vivo. Antimicrob Agents Chemother **55:**2891-2896.
509 23. **Sreevatsan S, Stockbauer KE, Pan X, Kreiswirth BN, Moghazeh SL, Jacobs WR,**
510   **Jr., Telenti A, Musser JM.** 1997. Ethambutol resistance in Mycobacterium
511   tuberculosis: critical role of embB mutations. Antimicrobial agents and chemotherapy
512   **41:**1677-1681.
513 24. **Bakuła Z, Napiórkowska A, Bielecki J, Augustynowicz-Kopeć E, Zwolska Z,**
514   **Jagielski T.** 2013. Mutations in the embB gene and their association with ethambutol
515   resistance in multidrug-resistant Mycobacterium tuberculosis clinical isolates from
516   Poland. Biomed Res Int **2013:**167954.
517 25. **Safi H, Fleischmann RD, Peterson SN, Jones MB, Jarrahi B, Alland D.** 2010.
518   Allelic exchange and mutant selection demonstrate that common clinical embCAB
519   gene mutations only modestly increase resistance to ethambutol in Mycobacterium
520   tuberculosis. Antimicrob Agents Chemother **54:**103-108.
521 26. **Hazbón MH, Brimacombe M, Bobadilla del Valle M, Cavatore M, Guerrero MI,**
522   **Varma-Basil M, Billman-Jacobe H, Lavender C, Fyfe J, García-García L, León**
523   **CI, Bose M, Chaves F, Murray M, Eisenach KD, Sifuentes-Osornio J, Cave MD,**
524   **Ponce de León A, Alland D.** 2006. Population genetics study of isoniazid resistance
525   mutations and evolution of multidrug-resistant Mycobacterium tuberculosis.
526   Antimicrob Agents Chemother **50:**2640-2649.
527 27. **Cavusoglu C, Karaca-Derici Y, Bilgic A.** 2004. In-vitro activity of rifabutin against
528   rifampicin-resistant Mycobacterium tuberculosis isolates with known rpoB mutations.
529   Clin Microbiol Infect **10:**662-665.
530 28. **Nelson MC, Morrison HG, Benjamino J, Grim SL, Graf J.** 2014. Analysis,
531   optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys.
532   PLoS One **9:**e94249.
533 29. **Achtman M.** 2008. Evolution, population structure, and phylogeography of
534   genetically monomorphic bacterial pathogens. Annu Rev Microbiol **62:**53-70.
535 30. **Al-Moamary MS, Black W, Bessuille E, Elwood RK, Vedal S.** 1999. The
536   significance of the persistent presence of acid-fast bacilli in sputum smears in
537   pulmonary tuberculosis. Chest **116:**726-731.
538 31. **Zhang X, Zhao B, Huang H, Zhu Y, Peng J, Dai G, Jiang G, Liu L, Zhao Y, Jin Q.**
539   2013. Co-occurrence of amikacin-resistant and -susceptible Mycobacterium
540   tuberculosis isolates in clinical samples from Beijing, China. J Antimicrob Chemother
541   **68:**1537-1542.
542 32. **Tolani MP, D'souza DT, Mistry NF.** 2012. Drug resistance mutations and
543   heteroresistance detected using the GenoType MTBDRplus assay and their
544   implication for treatment outcomes in patients from Mumbai, India. BMC Infect Dis
545   **12:**9.
546 33. **Cullen MM, Sam NE, Kanduma EG, McHugh TD, Gillespie SH.** 2006. Direct
547   detection of heteroresistance in Mycobacterium tuberculosis using molecular
548   techniques. J Med Microbiol **55:**1157-1158.
549 34. **Falagas ME, Makris GC, Dimopoulos G, Matthaiou DK.** 2008. Heteroresistance: a
550   concern of increasing clinical significance? Clin Microbiol Infect **14:**101-104.

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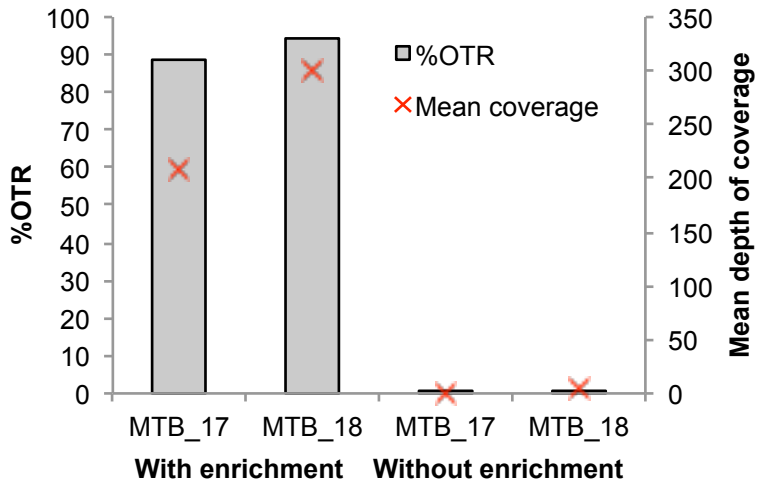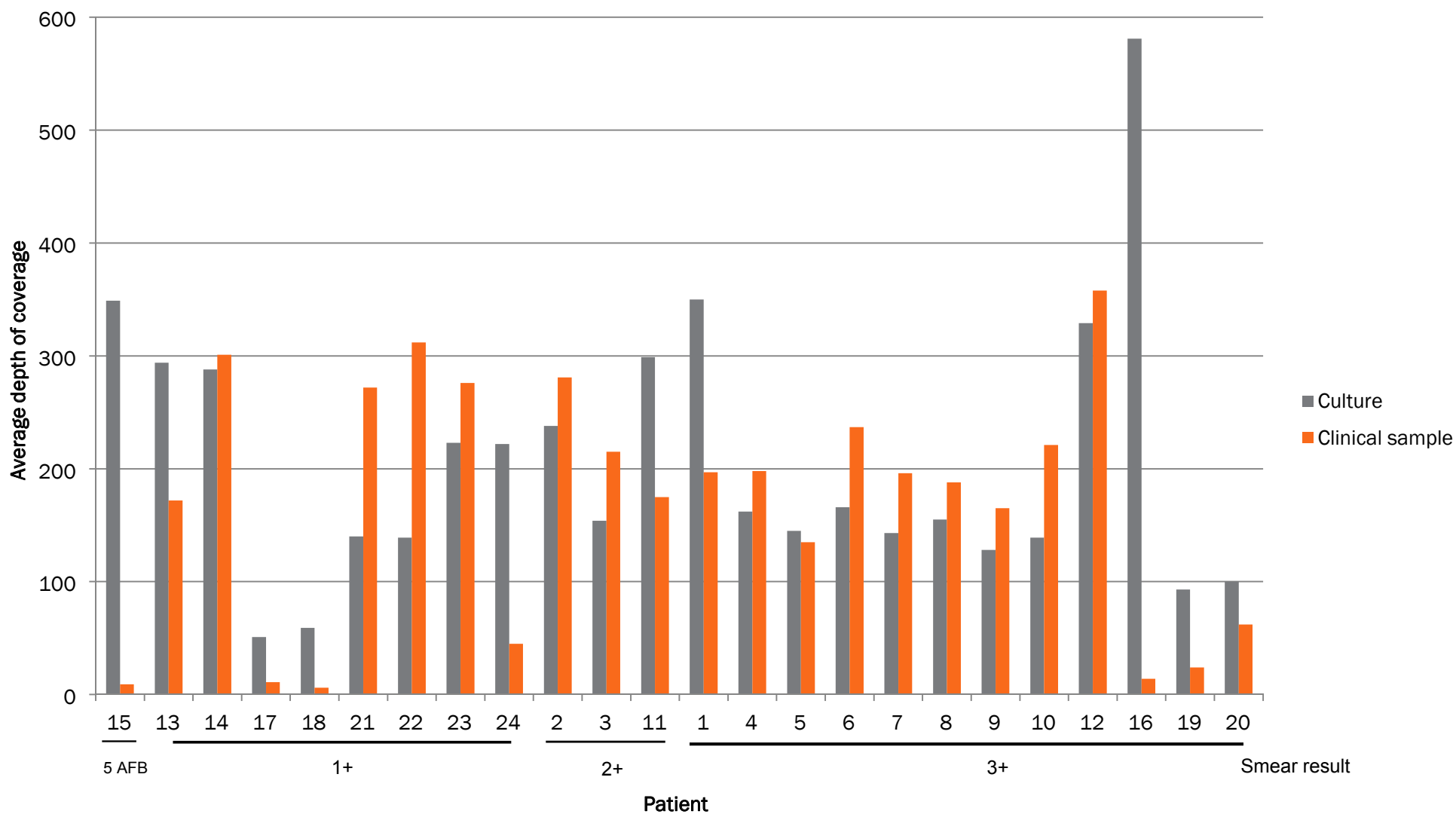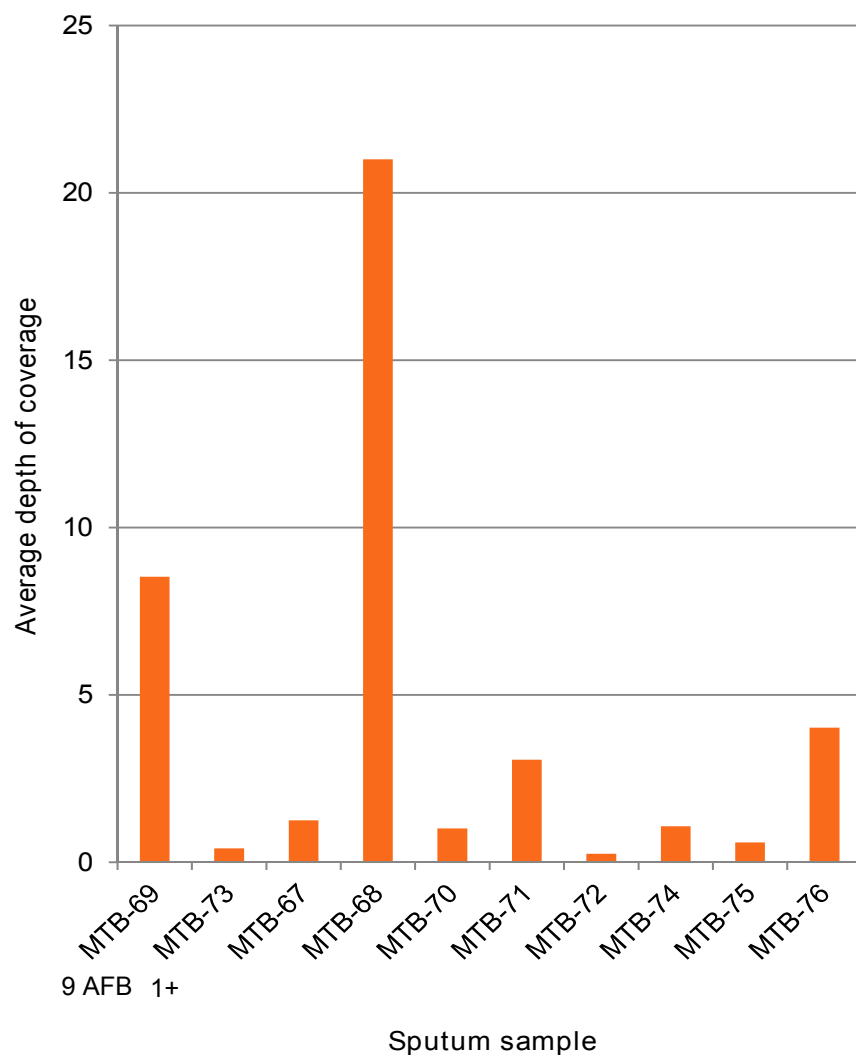554   sequencer. PeerJ **2:**e585.

555

556

557

Figure 1

Figure 2

Figure 3
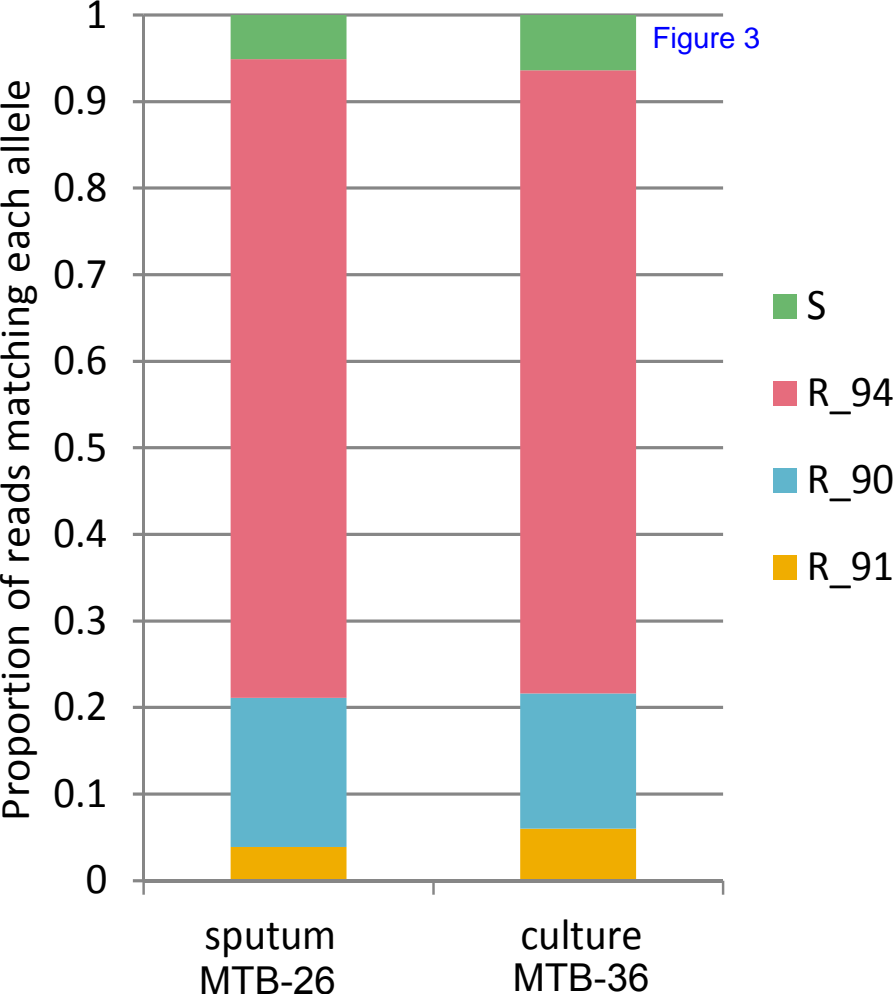
Table 1

| Patient	Sputum positivity	Sample	Type	Rif	Inh	Emb	Pza	Str*	Ofl*	Pas*	Amg*	Thi*
1	3+		Culture phenotype	S	S	S	S	NA	NA	NA	NA	NA
		MTB-27	Culture genotype									
		MTB-17	Sputum genotype									
2	2+		Culture phenotype	S	S	S	S	NA	NA	NA	NA	NA
		MTB-28	Culture genotype							R		
		MTB-18	Sputum genotype							R		
3	2+		Culture phenotype	S	S	S	S	NA	NA	NA	NA	NA
		MTB-29WE	Culture genotype									
		MTB-19	Sputum genotype									
4	3+		Culture phenotype	R	R	R	R	R	S	S	R (Kan)	S
		MTB-30WE	Culture genotype	R	R	Low R	R	R			R	
		MTB-20	Sputum genotype	R	R	Low R	R	R			R	
5	3+		Culture phenotype	S	R	S	R	R	R	S	R (Kan & Amk)	R
		MTB-31WE	Culture genotype	R	R	Low R		R	R		R	
		MTB-21	Sputum genotype	R	R	Low R		R	R		R (Kan)	
6	3+		Culture phenotype	R	R	S	R	R	R	S	R (Kan)	R
		MTB-32WE	Culture genotype	R	R	Low R	R	R	R		R	R
		MTB-22	Sputum genotype	R	R	Low R	R	R	R		R	R
7	3+		Culture phenotype	R	R	R	R	R	R	R	R (Cap)	R
		MTB-33WE	Culture genotype	R	R	R		R	R			
		MTB-23	Sputum genotype	R	R	R		R	R			
8	3+		Culture phenotype	R	R	R	R	R	R	R	S	S
		MTB-34WE	Culture genotype	R	R	Low R	R	R	R			
		MTB-24	Sputum genotype	R	R	Low R	R	R	R			
9	3+		Culture phenotype	R	R	R	NA	R	NA	NA	NA	NA
		MTB-35WE	Culture genotype	R	R	R	R	R				
		MTB-25	Sputum genotype	R	R	R	R	R				
10	3+		Culture phenotype	R	R	R	R	R	R	S	S	S
		MTB-36WE	Culture genotype	R	R	Low R	R	R	R			
		MTB-26	Sputum genotype	R	R	Low R	R	R	R			
11	2+		Culture phenotype	S	S	S	S	NA	NA	NA	NA	NA
		MTB-45	Culture genotype									
		MTB-37	Sputum genotype									
12	3+		Culture phenotype	S	R	S	S	NA	NA	NA	NA	NA
		MTB-46	Culture genotype		R							
		MTB-38	Sputum genotype		R							
13	1+		Culture phenotype	R	R	R	R	R	S	S	S	S
		MTB-47	Culture genotype	R	R	R	R	R				
		MTB-39	Sputum genotype	R	R	R	R	R				
14	1+		Culture phenotype	R	R	S	R	R	R	R	S	S
		MTB-48	Culture genotype	R	R	Low R	R	R				
		MTB-40	Sputum genotype	R	R	Low R	R	R				
15	5 AFB		Culture phenotype	R	R	S	R	R	R	NA	S	NA
		MTB-49	Culture genotype	R	R		R	R	R	R		
		MTB-41	Sputum genotype				Below level of detection					
16	3+		Culture phenotype	S	R	S	S	NA	NA	NA	NA	NA
		MTB-50	Culture genotype		R							
		MTB-42	Sputum genotype		R							
17	1+		Culture phenotype	R	R	S	S	NA	S	NA	S	NA
		MTB-51	Culture genotype	R	R	Low R				R		
		MTB-43	Sputum genotype				Below level of detection					
18	1+		Culture phenotype	S	S	S	S	NA	NA	NA	NA	NA
		MTB-52	Culture genotype					R				
		MTB-44	Sputum genotype					R				
19	3+		Culture phenotype	S	S	S	S	NA	NA	NA	NA	NA
		MTB-60	Culture genotype							R		
		MTB-53	Sputum genotype							R		
20	3+		Culture phenotype	S	S	S	S	NA	NA	NA	NA	NA
		MTB-61	Culture genotype									
		MTB-54	Sputum genotype									
21	1+		Culture phenotype	S	S	S	S	NA	NA	NA	NA	NA
		MTB-63	Culture genotype									
		MTB-55	Sputum genotype									
22	1+		Culture phenotype	S	S	S	S	NA	NA	NA	NA	NA
		MTB-64	Culture genotype									
		MTB-56	Sputum genotype									
23	1+		Culture phenotype	S	S	S	S	NA	NA	NA	NA	NA
		MTB-65	Culture genotype									
		MTB-57	Sputum genotype									
24	1+		Culture phenotype	R	R	S	S	NA	S	NA	S	S
		MTB-66	Culture genotype	R	R	Low R				R		
		MTB-59	Sputum genotype	R	R	Low R				R		

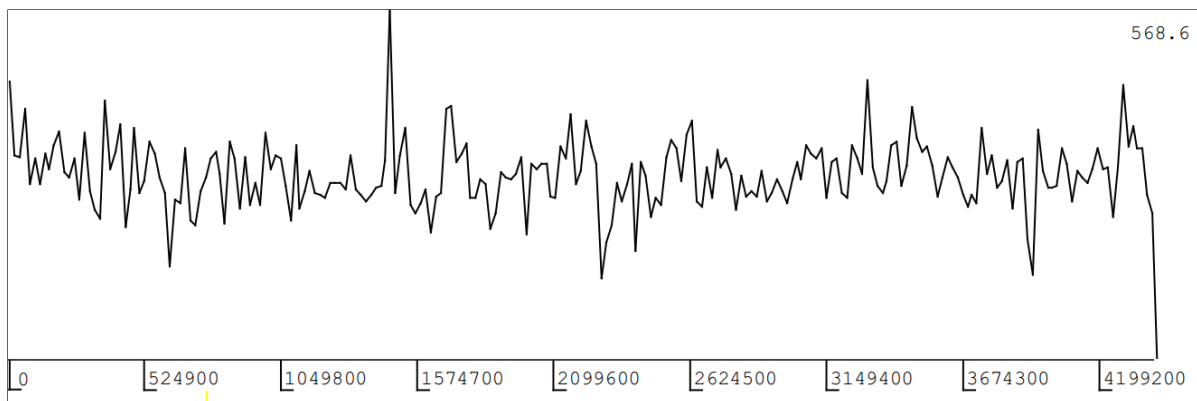**Supplementary table 1**: Available details of samples sequenced in this study.

Excel spreadsheet

Supplementary Figure 1: Flow diagram for preparation of M. tb samples for enrichment based WGS from receipt of extracted genomic DNA to final data report. Times given are based on 16 samples processed manually or 96 samples processed using automation; Enrichment- we have seen comparable data from 16h vs 24h enrichment; MiSeq times are depending on MiSeq cartridge and chemistry used (2x75bp v3 run = ~16h, 2x300bp v3 run = ~68h)
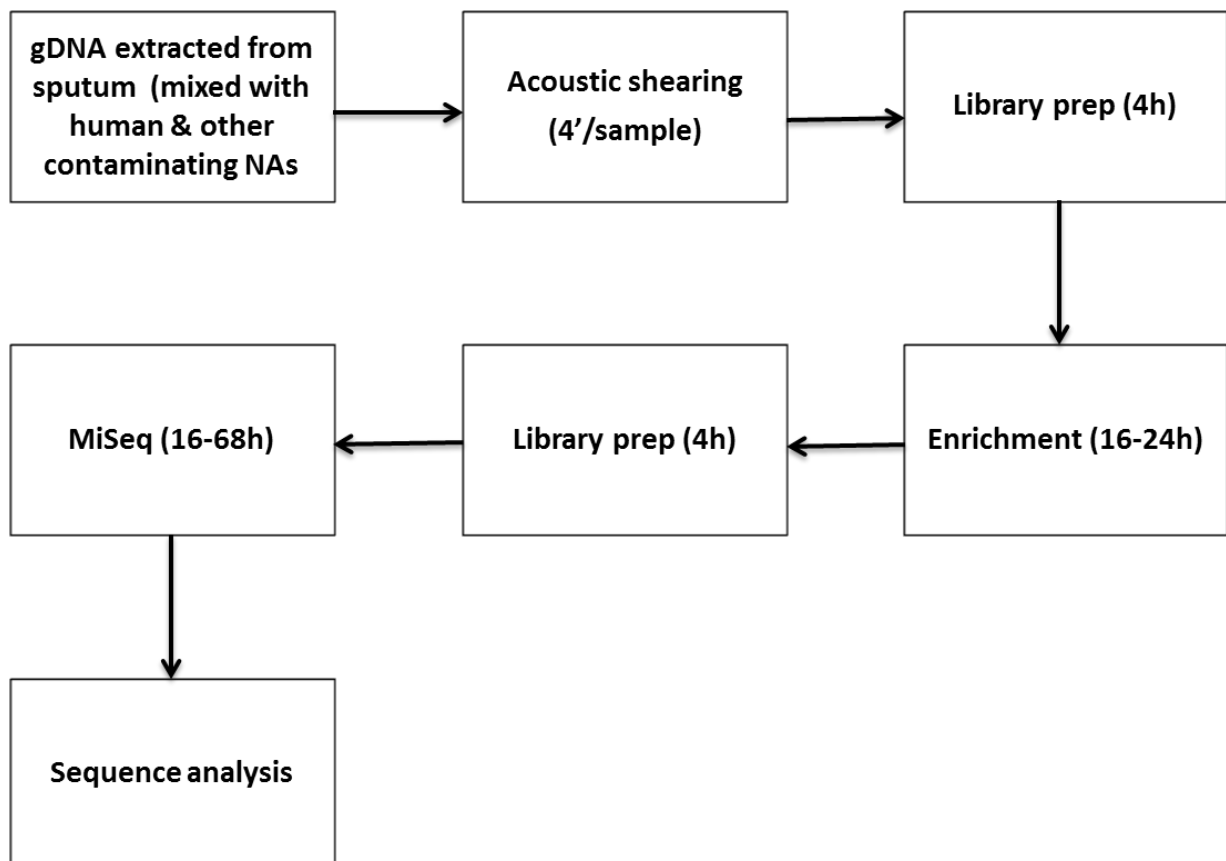
Supplementary Figure 2: Comparison of sequence results from culture with and without enrichment for two samples.
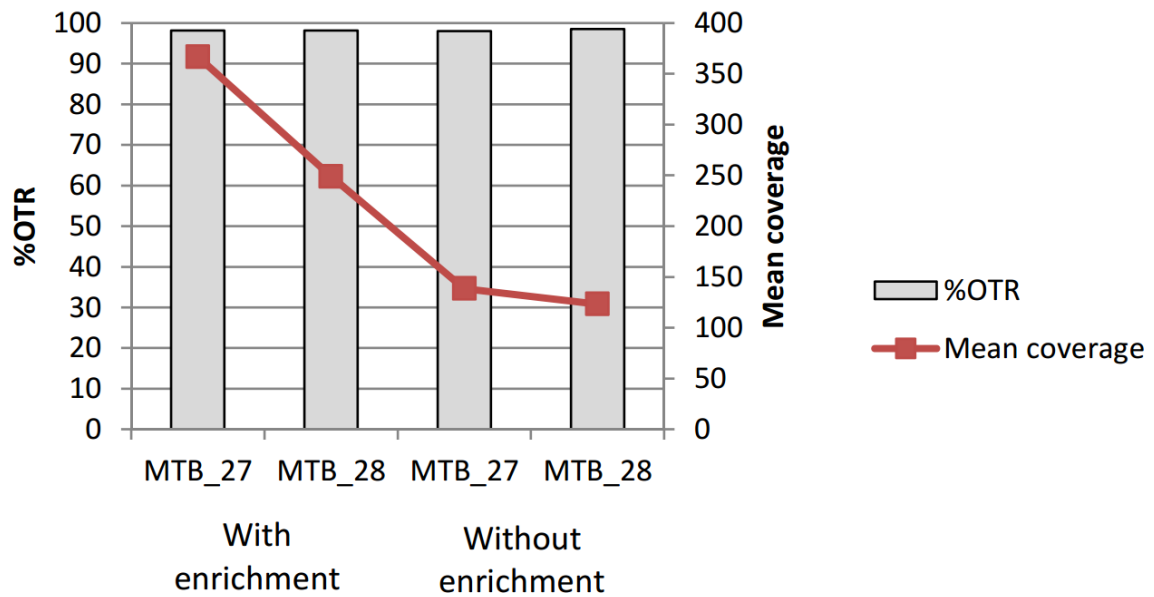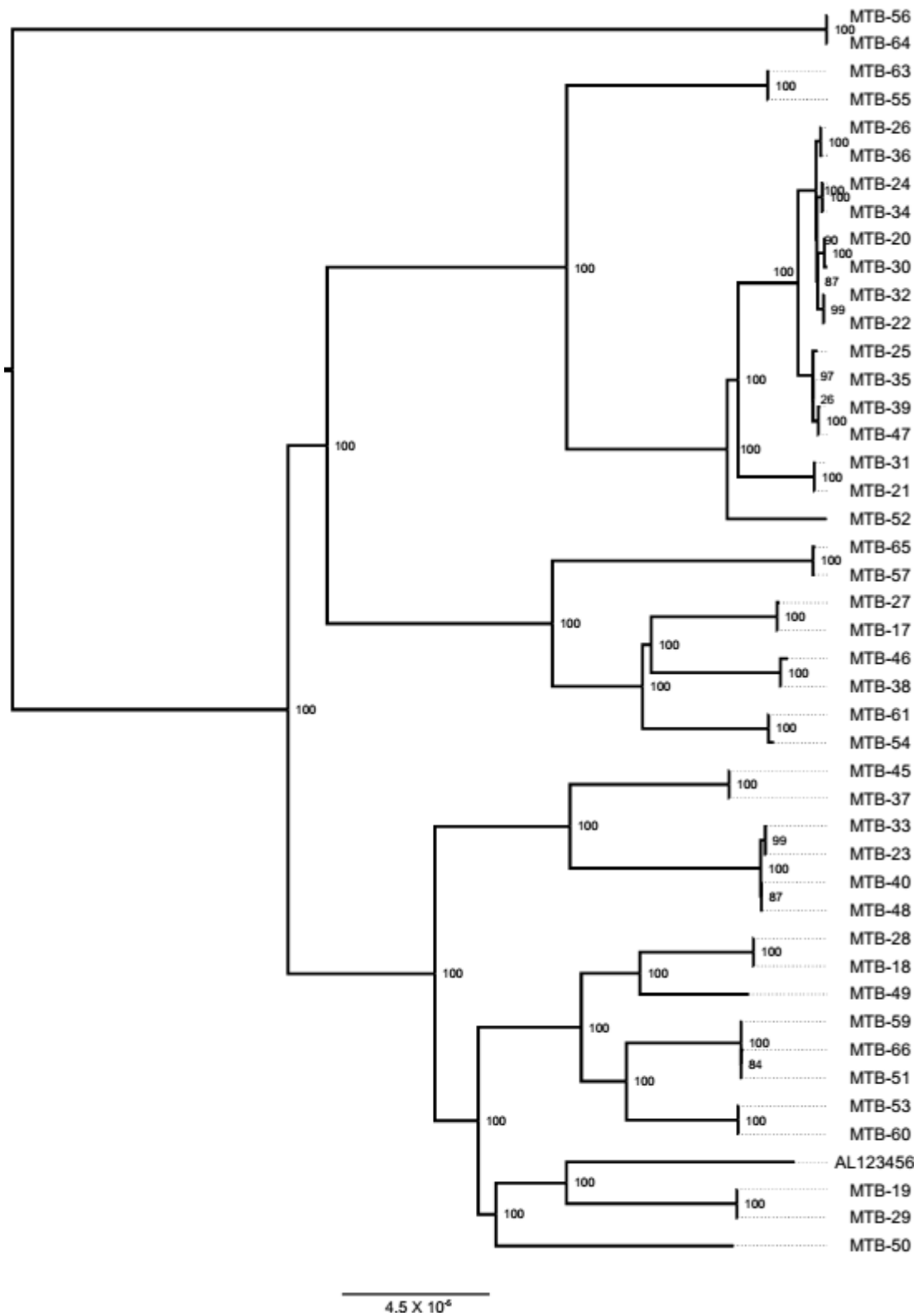

Supplementary Figure 3: Coverage plot of enriched sputum sample MTB-40 mapped against H37Rv. Maximum point on graph is 568.6 and minimum is 0. X axis indicates base position along the genome. Adapted from image generated using Artemis (Carver et. al. 2012 http://www.sanger.ac.uk/resources/software/artemis/).

Supplementary Figure 4: Maximum likelihood tree of all samples where at least 1x depth of coverage was obtained. Clade support is indicated by number of bootstrap replicates (out of 100). Scale bar represents substitutions per site. Some samples were identified as being on long branches or positioned close to nodes. Further investigation revealed that they all had at least 10% missing data in the SNP alignment, due to low coverage or heterozygosity, so could not be placed accurately on the tree. The samples indicated in red were removed from the final maximum likelihood tree shown in Supplementary Figure 3.

Supplementary Figure 5: Maximum likelihood tree of 45 samples that had high coverage and could be accurately placed on the tree (see Supplementary Figure 2). Clade support is indicated by number of bootstrap replicates (out of 100). The tree was constructed using RAxML. Scale bar represents substitutions per site.

Supplementary Figure 6: Deletion of pncA and surrounding genes identified in patient 7. Coverage plot is shown in grey.
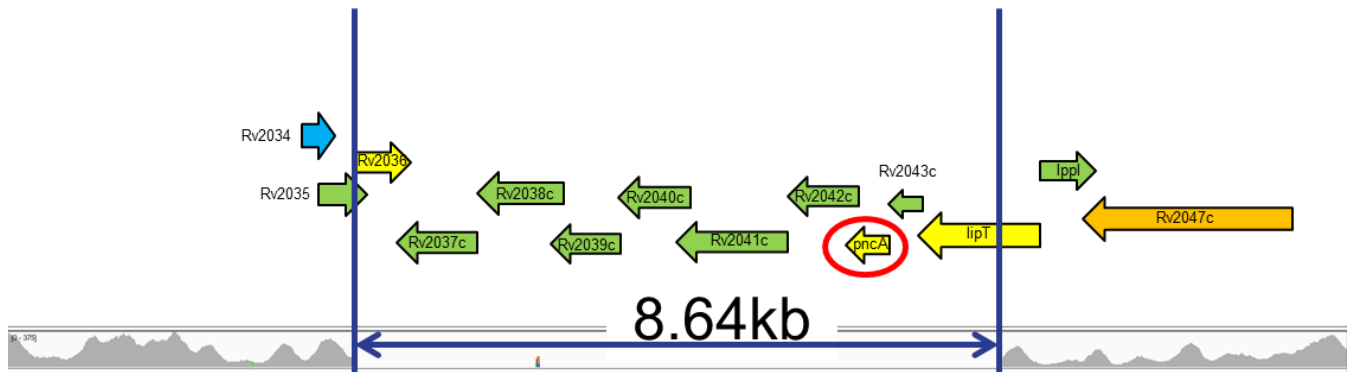
Supplementary table 2: Resistance genotypes identified in this study that passed all quality criteria and were found at greater than 10% frequency. The position of the mutation identified in column D refers to H37Rv (AL123456.3). Support on the forward, reverse and from individual nucelotides is shown. Any variants that didn't match the observed phenotype were cross-references against the literature, and are discussed in the main text.

Excel spreadsheet