

1   **Title:** Same-day diagnostic and surveillance data for tuberculosis via whole  
2   genome sequencing of direct respiratory samples.  
3  
4

5   **Authors**

6   Antonina A. Votintseva<sup>1#\*</sup>, Phelim Bradley<sup>2\*</sup>, Louise Pankhurst<sup>1\*</sup>, Carlos del Ojo  
7   Elias<sup>2</sup>, Matthew Loose<sup>3</sup>, Kayzad Nilgiriwala<sup>4</sup>, Anirvan Chatterjee<sup>4</sup>, E. Grace Smith<sup>5</sup>,  
8   Nicolas Sanderson<sup>1</sup>, Timothy M. Walker<sup>1</sup>, Marcus R. Morgan<sup>6</sup>, David H. Wyllie<sup>1,7</sup>,  
9   A. Sarah Walker<sup>1,8</sup>, Tim E. A. Peto<sup>1,8</sup>, Derrick W. Crook<sup>1,8+</sup>, Zamin Iqbal<sup>2+\*</sup>  
10  
11

12   **Authors' affiliations**

13   <sup>1</sup>Nuffield Department of Clinical Medicine, University of Oxford, John Radcliffe  
14   Hospital, Oxford, OX3 9DU, United Kingdom

15   <sup>2</sup>Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3  
16   7BN, UK

17   <sup>3</sup>School of Life Sciences, University of Nottingham, Nottingham, NG7 2UH

18   <sup>4</sup>Foundation for Medical Research, Mumbai

19   <sup>5</sup>Regional Centre for Mycobacteriology, PHE Public Health Laboratory  
20   Birmingham. Heartlands Hospital, Birmingham B9 5SS, UK

21   <sup>6</sup>Microbiology Laboratory, John Radcliffe Hospital, Oxford University Hospitals  
22   NHS Trust, Oxford, OX3 9DU, United Kingdom

23   <sup>7</sup>The Jenner Institute, University of Oxford, Roosevelt Drive, Oxford OX3 7DQ,  
24   United Kingdom

25   <sup>8</sup>National Institute for Health Research (NIHR) Oxford Biomedical Research  
26   Centre, John Radcliffe Hospital, Oxford, OX3 9DU, United Kingdom  
27  
28  
29  
30  
31

32   + These authors contributed equally

33   ★ These authors contributed equally

34   #Corresponding authors:

35   Dr Antonina A. Votintseva; Address: Nuffield Department of Clinical Medicine,  
36   University of Oxford, John Radcliffe Hospital, Level 7, Oxford, OX3 9DU, United  
37   Kingdom<sup>1</sup>. Email: a.votintseva@gmail.com  
38

39   Dr Zamin Iqbal; Address: Wellcome Trust Centre for Human Genetics, Roosevelt  
40   Drive, Oxford, OX3 7BN.

41   Email: [zam@well.ox.ac.uk](mailto:zam@well.ox.ac.uk)

42   Running title: Same-day TB WGS from direct respiratory samples  
43  
44  
45  
46  
47  
48  
49

50

## 51 Abstract

52

53 Routine full characterization of *Mycobacterium tuberculosis* (TB) is culture-based, taking many weeks. Whole-genome sequencing (WGS) can generate  
54 antibiotic susceptibility profiles to inform treatment, augmented with strain  
55 information for global surveillance; such data could be transformative if  
56 provided at or near point of care.

57

58 We demonstrate a low-cost DNA extraction method for TB WGS direct from  
59 patient samples. We initially evaluated the method using the Illumina MiSeq  
60 sequencer (40 smear-positive respiratory samples, obtained after routine clinical  
61 testing, and 27 matched liquid cultures). *M. tuberculosis* was identified in all 39  
62 samples from which DNA was successfully extracted. Sufficient data for  
63 antibiotic susceptibility prediction was obtained from 24 (62%) samples; all  
64 results were concordant with reference laboratory phenotypes. Phylogenetic  
65 placement was concordant between direct and cultured samples. Using an  
66 Illumina MiSeq/MiniSeq the workflow from patient sample to results can be  
67 completed in 44/16 hours at a cost of £96/£198 per sample.

68

69 We then employed a non-specific PCR-based library preparation method for  
70 sequencing on an Oxford Nanopore Technologies MinION sequencer. We applied  
71 this to cultured *Mycobacterium bovis* BCG strain (BCG), and to combined culture-  
72 negative sputum DNA and BCG DNA. For the latest flowcell, the estimated  
73 turnaround time from patient to identification of BCG was 6 hours, with full  
74 susceptibility and surveillance results 2 hours later. Antibiotic susceptibility  
75 predictions were fully concordant. A critical advantage of the MinION is the  
76 ability to continue sequencing until sufficient coverage is obtained, providing a  
77 potential solution to the problem of variable amounts of *M. tuberculosis* in direct  
78 samples.

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

## 98      **Introduction**

99  
100 The long-standing gold standard for *Mycobacterium tuberculosis* drug  
101 susceptibility testing (DST) is the phenotypic culture-based approach, which is  
102 time-consuming and laborious. First-line tuberculosis (TB) treatment includes  
103 four drugs (rifampicin, isoniazid, ethambutol and pyrazinamide) but with the  
104 spread of multi-drug resistant strains, there is a growing need for data on  
105 second-line drugs, including the fluoroquinolones, and aminoglycosides.  
106  
107 Due to long turnaround times for phenotypic testing (up to two months), these  
108 are often preceded by WHO-endorsed molecular methods such as the GenoType  
109 MTBDRplus and MTBDRsl assays (Hain Lifescience GmbH, Germany), and Xpert  
110 MTB/RIF (Cepheid, USA). These potentially culture-free, PCR-based tests rapidly  
111 identify species and detect the most common drug resistance conferring  
112 mutations. However, this technology is limited by the number of mutations that  
113 can be probed. This limitation is of concern, given the many low frequency drug  
114 resistance conferring mutations in *M. tuberculosis*, particularly for second-line  
115 drugs (1). Consistent with this concern, the proportion of phenotypically  
116 resistant samples which are detectable by MTBDRplus range from 21-25% for  
117 the second-line drugs capreomycin and kanamycin (2) to 98.4% and 91.4% for  
118 the critical first-line drugs rifampicin and isoniazid (3). A potential solution is to  
119 sequence amplicons targeting a wider range of resistance conferring genes, as  
120 previously demonstrated (4).  
121  
122 The potential of whole genome sequencing (WGS) as a diagnostic assay has been  
123 repeatedly identified (5-7). Recent studies based on WGS of mycobacteria have  
124 evaluated WGS-based susceptibility predictions (1, 8-10), species identification,  
125 and elucidation of epidemiology (11-16). This has culminated in the first  
126 successful application of WGS as a clinical diagnostic for mycobacteria from early  
127 positive liquid cultures (16). Moreover, WGS was performed at a cost  
128 comparable with existing phenotypic assays and offered faster turnaround times.  
129  
130 Generating WGS information directly from patient samples, and avoiding the  
131 need for culture, would be transformative. However, direct samples contain  
132 highly variable amounts of mycobacterial cells mixed with other bacterial and  
133 human cells; the latter accounting for up to 99.9% of DNA present. Furthermore,  
134 mycobacterial cells may aggregate due to the high mucus content of some  
135 samples; meaning sample volume and Acid Fast Bacillus (AFB) count may not  
136 represent the total quantity of mycobacteria available. Direct samples therefore  
137 require pre-processing to homogenize and enrich for mycobacteria by depleting  
138 other cells/DNA. The challenges of direct sample processing were illustrated by  
139 two studies assessing the feasibility of WGS directly from clinical samples (17-  
140 18). By sequencing eight smear-positive sputum samples subjected to  
141 differential lysis followed by DNA extraction with a commercial kit, Doughty and  
142 colleagues were able to achieve only 0.002-0.7X depth of coverage for *M.*  
143 *tuberculosis* with 20.3-99.3% of sequences mapping to the human genome. 7/8  
144 samples could be assigned to *M. tuberculosis* complex, but none had sufficient  
145 data for drug susceptibility prediction. In a second study, Brown and colleagues  
146 applied a SureSelect target enrichment method (Agilent, USA) to capture *M.*

147 *tuberculosis* DNA prior to WGS. 20/24 smear-positive samples achieved 90%  
148 genome coverage with  $\geq 20x$  depth; sufficient for prediction of species and  
149 antibiotic susceptibility. However, the protocol was slow (2-3 days) and may be  
150 prohibitively expensive for use in low-income settings.

151  
152 In this study we test a simple low-cost DNA extraction method using Illumina  
153 MiSeq WGS on 40 smear-positive, primary respiratory samples from *M.*  
154 *tuberculosis* infected patients. We evaluate the protocol in terms of DNA  
155 obtained, species assignment of the sequenced reads, and our ability to obtain  
156 key clinical data (detection of *M. tuberculosis* and antibiotic susceptibility  
157 prediction) along with epidemiological information (placement on phylogenetic  
158 tree). These data would enable a single test to deliver the core information for  
159 both patient and public health in <48 hours using Illumina-based WGS. We also  
160 develop an approach for WGS using the highly portable, random-access, Oxford  
161 Nanopore Technologies (ONT) MinION, reducing potential turnaround time to 8  
162 hours.

163  
164  
165  
166  
167  
168

## 169 Materials and Methods

170  
171

### 172 Sample selection and processing

173 Direct respiratory Ziehl-Neelsen (ZN)-positive samples with acid-fast bacilli  
174 (AFB) scorings from +1 to +3 were collected from patients with subsequently  
175 confirmed *M. tuberculosis* infections at the John Radcliffe Hospital, Oxford  
176 Universities NHS Foundation Trust, Oxford, UK (n=18), and Birmingham  
177 Heartlands Hospital NHS Foundation Trust, Birmingham, UK (n=22). 2/18  
178 Oxford samples were culture negative specimens taken 2.5 months apart from  
179 the same patient undergoing treatment for *M. tuberculosis*. If available,  
180 corresponding Mycobacterial Growth Indicator Tube (MGIT) cultures were  
181 collected for each direct sample (Oxford n=11, Birmingham n=17). Two ZN and  
182 culture negative direct respiratory samples were also collected from the John  
183 Radcliffe Hospital.

184  
185 The discarded direct samples were collected only after sufficient material had  
186 been obtained for the routine diagnostic workflow, including the requirement to  
187 ensure that enough sample volume remained if re-culture was requested.  
188 Consequently, our samples were of lower volume and quality than would be the  
189 case if the method were used routinely. While waiting for the routine laboratory  
190 results, samples were stored at +4C and later processed in batches of 5-12. All  
191 ZN-positive samples were digested and decontaminated with NAC-PAC RED kit  
192 (AlphaTec, USA). Direct samples and corresponding MGIT culture aliquots (1  
193 mL) were heat inactivated in a thermal block after sonication (20 min, 35 kHz)  
194 for 30 min and 2 h at 95C, respectively. MGITs were inactivated for 2 hours

195 owing to their high bacterial load. Before DNA extraction samples were stored at  
196 +4C.  
197

#### 198 **DNA extraction and Illumina MiSeq sequencing**

199 Mycobacterial DNA from MGIT cultures was extracted using a previously  
200 validated ethanol precipitation method (19). DNA from ZN-positive direct  
201 samples was extracted using a modified version of this protocol. These  
202 modifications included a saline wash followed by MolYsis Basic5 kit (Molzym,  
203 Germany) treatment for the removal of human DNA, and addition of GlycoBlue  
204 co-precipitant (LifeTechnologies, USA) to the ethanol precipitation step  
205 (Supplementary Figure 1).

206 Libraries were prepared for the MiSeq Illumina sequencing using a modified  
207 Illumina Nextera XT protocol (19). Samples were sequenced using the MiSeq  
208 Reagent Kit v2, 2 x 150bp in batches of 9-12 per flow-cell.  
209

#### 210 **DNA extraction for ONT MinION and Illumina MiniSeq sequencing**

211 ZN/culture-negative sputum and BCG (Pasteur strain; cultivated at 37C in MGIT  
212 tubes) DNA was extracted using a modified version of that in (19). Briefly,  
213 following a saline wash, samples were re-suspended in 100 µL of molecular  
214 grade water and subjected to three rounds of bead-beating at 6 m/s for 40  
215 seconds. The beads were pelleted by centrifugation at 16,100 xg for 10 minutes  
216 and 50 µL supernatant cleaned using 1.8x volume AMPure beads (Beckman  
217 Coulter, UK). Samples were eluted in 25 µL molecular grade water, and  
218 quantified using the Qubit fluorimeter (Thermo Fisher Scientific, USA). (Steps I,  
219 III, V and VI of Miseq protocol, Supplementary Figure 1.)  
220

#### 221 **MiniSeq sequencing**

222 Extracted ZN-negative sputum DNA and pure BCG DNA were combined in a  
223 50:50 ratio (0.5 ng each) and libraries prepared alongside pure BCG DNA (1 ng)  
224 using a modified Illumina Nextera XT protocol (19). BCG and two BCG+sputum  
225 DNA samples were sequenced at Illumina Cambridge Ltd. UK, using a Mid Output  
226 kit (FC-420-1004) reading 15 tiles and with 101 cycles.  
227

#### 228 **MinION sequencing**

229 All MinION sequencing utilized the best sample preparation kits and flow cells  
230 available at the time. A single ZN-negative sputum extract was divided into three  
231 equal concentration aliquots (187 ng), and BCG DNA added at 5%, 10% and 15%  
232 of the total sputum DNA concentration. These 5-15% spikes represent the lower  
233 end of the spectrum seen in the MiSeq samples above (see Figure 2a). These  
234 samples, along with pure BCG DNA, were prepared following ONTs PCR-based  
235 protocol for low-input libraries (DP006\_revB\_14Aug2015), using modified  
236 primers supplied by ONT, a 20 ng DNA input into the PCR reaction, and LongAmp  
237 Taq 2X Master Mix (New England Biolabs, USA). PCR conditions were as follows:  
238 initial denaturation at 95°C for 3 minutes, followed by 18 cycles of 95°C for 15s,  
239 62°C for 15s, and 65°C for 2.5 minutes, and a final extension at 65°C for 5  
240 minutes. Samples were cleaned in 0.4x volume AMPure beads and the PCR  
241 product assessed using the Qubit fluorimeter and TapeStation (Agilent, UK). The  
242

244 final elution was into 10 µL 50 mM NaCl, 10 mM Tris.HCl pH8.0. Finally, 1 µL of  
245 PCR-Rapid Adapter (PCR-RAD; supplied by ONT) was added and samples  
246 incubated for 5 minutes at room temperature to generate pre-sequencing mix.  
247 The pre-sequencing mix was prepared for loading onto flow cells following  
248 standard ONT protocols, with a loading concentration of 50 – 100 fmol.  
249  
250 Using the 15% BCG spiked sputum DNA prepared above, amplification was  
251 repeated using Phusion High-Fidelity PCR Master Mix with DMSO (New England  
252 BioLabs, USA). Gradient PCR was performed to identify the optimal annealing  
253 temperature for recovery of BCG DNA (data not shown). Final PCR conditions  
254 were as follows: initial denaturation at 98°C for 30s, followed by 18 cycles of  
255 98°C for 10s, 59°C for 15s, and 72°C for 1.5 minutes, and a final extension of 72°C  
256 for 10 minutes. Following PCR, the sample was prepared for sequencing as  
257 described above. The final loading concentration was approximately 27 fmol.  
258  
259 The above samples were sequenced using R9 spot-on generation flow cells and  
260 the 48-hour protocol for FLO-MIN105 (ONT, UK). Base calling was performed via  
261 the Metrichor EPI2ME service (ONT, UK) using the 1D RNN for SQK-RAD001  
262 v1.107 workflow.  
263  
264 Subsequently, a new 15% BCG spiked sputum was prepared as described above  
265 using Phusion Master Mix with DMSO. Sequencing was performed using R9.4  
266 spot-on generation flow cells and the 48-hour FLO-MIN106 protocol (ONT, UK).  
267 Final loading concentration was 43 fmol. Base calling was performed after  
268 sequencing was complete using Albacore (ONT, UK), as base calling via Metrichor  
269 failed.  
270  
271  
272  
273

#### 274 **Bioinformatic analysis of Illumina data**

275 To determine levels of contamination and *M. tuberculosis* in samples, reads were  
276 immediately mapped using bwa\_mem (20) to the human reference genome  
277 GRCh37 (hg19) and human reads counted and permanently discarded.  
278 Remaining stored reads were then mapped to the *M. tuberculosis* H37Rv  
279 reference strain (GenBank NC\_018143.2), and any unmapped reads were then  
280 mapped to nasal, oral and mouth flora available in the NIH Human Microbiome  
281 Project database (<http://www.hmpdacc.org/>). A minimum reference genome  
282 coverage depth of 5 was required for phylogenetic analysis to proceed.  
283

284 Mycobacterial species and antibiotic resistance to isoniazid, rifampicin,  
285 ethambutol, pyrazinamide, streptomycin, aminoglycosides (including  
286 capreomycin, amikacin and kanamycin) and fluoroquinolones (including  
287 moxifloxacin, ofloxacin, and ciprofloxacin) was predicted using Mykrobe  
288 predictor software (21) v0.3.5 updated with a new validated catalogue of  
289 resistance conferring genetic mutations (1). For samples where the estimated  
290 depth of kmer-coverage of *M. tuberculosis* reported by Mykrobe predictor fell  
291 below 3x, no resistance predictions were made. The precise command used

292 was; ``mykrobe predict SAMPLE\_ID tb -1 FASTQ –panel walker-2015 –min-depth  
293 3".  
294  
295

### 296 **Phylogenetic analysis**

297 Conservative SNP calls were made using Cortex (22) (independent workflow,  
298 k=31) on 3480 samples from (1). Singleton variants were discarded, and a de-  
299 duplicated list of 68695 SNPs was constructed. All samples (from our study and  
300 from (1)) were genotyped at these sites using the Cortex genotyping model (22).  
301 We then measured the number of SNP differences between paired direct and  
302 MGIT samples, counting only sites where both genotypes had high confidence  
303 (difference between log likelihood of called genotype and of uncalled genotype  
304 greater than 1), and neither site was called as heterozygous.  
305

306 Samples were placed on the phylogenetic tree of 3480 samples from (1) by  
307 identifying the leaf with the fewest SNP differences, across the 68695 sites.  
308 Placement therefore returns a closest leaf, and a SNP distance.  
309

### 310 **Statistical analysis**

311 Univariable and multivariable linear regression was used to identify  
312 independent factors affecting log10 DNA concentration after extraction. Analyses  
313 were performed using Stata 14.1 (2015, StataCorp, USA).  
314

### 315 **Bioinformatic Analysis of MinION Data**

316 Mykrobe predictor version v0.3.5-0-gd724461 was used to predict resistance  
317 from the MinION basecalled reads (command: mykrobe predict SAMPLE\_ID tb -1  
318 FASTQ –panel walker-2015 –min-depth 3).  
319

320 As for Illumina data, kmer coverage was estimated using a set of species-  
321 informative sequence probes defined in (21), and susceptibility predictions were  
322 only made if the median coverage was >3. Yield and timing was analysed using  
323 Poretools (23). For the R9.4 sample, Mykrobe predictor was applied to the  
324 cumulative read output at each hour. We excluded one false positive rifampicin  
325 resistance call seen at low coverage (the first 5 hours of sequencing) that was  
326 due to a known software bug (see Supplementary Text 1). Yield of BCG was  
327 measured by mapping to a BCG reference (accession BX248333.1).  
328

329 Phylogenetic placement of the 15% spike BCG sample sequenced on MinION R9.4  
330 was achieved as for the Illumina data - by genotyping 68695 SNPs using the  
331 Cortex model, and choosing the leaf with the fewest SNP differences across those  
332 sites.  
333

### 334 **MinION error analysis**

335 Error bias in the consensus of MinION R9 1D pure BCG reads was measured in  
336 two ways, using reads from the pure BCG sequencing run mentioned above.  
337

- 338 1. Reads were mapped to the *M. tuberculosis* reference genome using  
339 bwa\_mem, and then this was passed to the consensus tool racon (24). The

341 output of this was compared with the BCG reference genome using  
342 MUMMER (25). Any observed SNPs were considered to be errors, and  
343 bias in these errors was observed by looking at isolated SNPs (avoiding  
344 alignment artefacts due to nearby indels). The results are shown in  
345 Supplementary Table 1.

346 2. A *de novo* assembly was performed with Canu (26), and then this was  
347 compared with the BCG reference genome using MUMMER, as above. Full  
348 results shown in Supplementary Table 2.

349

350 The results were broadly consistent between the two methods. The mapping  
351 approach (number 1 above) found 28% of consensus errors were A->G and 60%  
352 were T->C. (Note these refer to the SNP with respect to the reference, not to  
353 errors within a single read.) The de novo assembly approach found 50% of  
354 consensus errors were A->G, and 44% were T->C.

355

356

### 357 Costing analysis

358 Basic costing included reagents required for sample decontamination, DNA  
359 extraction, MiSeq and Nanopore library preparations, and sequencing; correct as  
360 of November 2016. Generic laboratory consumables (e.g. pipette tips, tubes)  
361 were not included. SureSelect (Agilent, UK) costs, as used by Brown *et al.* (18),  
362 were obtained via a company representative and were correct of June 2016.  
363 United States Dollars (USD) were converted to Great British Pounds (GBP) at  
364 \$1.25 USD per GBP. See Supplementary Table 3 for details.

365

### 366 Ethics

367 For this study no ethical review was required because it was a laboratory  
368 methods development study focusing on bacterial DNA extracted from discarded  
369 samples identified only by laboratory numbers with no personal or clinical data.  
370 Sequencing reads identified as human based on fast mapping with BWA were  
371 counted and immediately permanently discarded (i.e. never stored  
372 electronically).

373

### 374 Accession numbers

375 Genome sequence data are in the process of being deposited to the Sequence  
376 Read Archive (SRA), NCBI. The MiSeq data is submitted under the study  
377 accession number SRP093599. The MiniSeq and MinION data accession numbers  
378 will be entered here when available.

379

380

## 381 Results

382

383

### 384 DNA extraction protocol and evaluation of Illumina sequencing output

385

386 DNA was extracted from 40 ZN-positive direct respiratory samples, of which 38  
387 were culture-confirmed *M. tuberculosis* ("culture-positive") and 2 were culture-  
388 negative. DNA was also extracted from 28 available corresponding MGIT  
389 cultures. All direct samples were the remainder of specimens available after

390 processing by the routine laboratory, and therefore had variable volume (median  
391 1.5 ml, IQR 0.5-3.1, range 0.25-15) and age (median 30 days from collection to  
392 processing, IQR 15-45, range 0-67). Most direct samples (78%; 31/40) could  
393 therefore be considered suboptimal on the basis of either low volume ( $\leq 1\text{ml}$ ) or  
394 long storage time ( $\geq 30$  days) or both.  
395

396 After DNA extraction, 33/40 (83%) direct samples and all 28 MGIT cultures  
397 yielded  $\geq 0.2 \text{ ng}/\mu\text{l}$  DNA, the amount recommended for MiSeq Illumina library  
398 preparation (Figure 1).  
399

400 There was no evidence that DNA yield was affected (either in multivariable or  
401 univariable models) by (1) sample type (sputum or bronchoalveolar lavage)  
402 ( $p=0.94$ ; univariable linear regression), (2) AFB scorings (from +1 to +3)  
403 ( $p=0.37$ ), (3) storage time prior to DNA extraction (days from collection)  
404 ( $p=0.51$ ) and (4) sample volume ( $p=0.28$ ). Although DNA concentration was  
405 measured and recorded after extraction, further data on DNA quality (e.g. DNA  
406 Integrity Number (DIN) provided by TapeStation (Agilent, USA)) were not  
407 routinely recorded.  
408

409 In total 39/40 direct samples with detectable DNA (37 culture-positive, 2  
410 culture-negative) and 27/28 MGIT cultures were sequenced on an Illumina  
411 MiSeq. One MGIT culture was not sequenced because the corresponding direct  
412 sample failed to yield measurable DNA. We used a lower than recommended  
413 DNA concentration threshold for MiSeq library preparation ( $>0.05 \text{ ng}/\mu\text{l}$  rather  
414 than  $>0.2 \text{ ng}/\mu\text{l}$ ) on the basis of previous experience of sequencing mycobacterial  
415 cultures with suboptimal amounts of DNA (19). 6/40 (15%) samples yielded  
416 DNA below the  $0.2 \text{ ng}/\mu\text{l}$  threshold. All sequenced direct samples produced  $\geq 1.5$   
417 million reads (median 3.6 million, IQR 2.9-5.0, range 1.5-12), as did all MGIT  
418 cultures (median 3.1 million, IQR 2.8-3.3, range 2.0-4.1).  
419  
420

## 421 Contamination levels of direct and MGIT samples

422

423 We assigned reads to categories *M. tuberculosis*, human, naso-pharyngeal flora  
424 (NPF) and “other” by mapping (see Methods). 77% (30/39) of direct samples  
425 contained  $<10\%$  human reads. However, only 46% (18/39) contained  $<10\%$  NPF  
426 and other bacterial reads, and 26% (10/39) contained  $>40\%$  of reads from non-  
427 mycobacterial, non-NPF, bacteria (Figure 2a). By comparison, MGIT culture  
428 samples showed much less contamination, as expected. (Figure 2b).  
429

## 430 Recovery of *M. tuberculosis* genome

431

432 Figure 3a shows the distribution of the *M. tuberculosis* reference genome depth  
433 of coverage across direct samples. Samples either have more than 10x depth and  
434 recover more than 90% of the genome, or have  $<3x$  depth and recover less than  
435 12% of the genome. The vertical dotted line delineates our threshold of 3x  
436 coverage, below which resistance predictions were not made. Figure 3b shows  
437 the amount of contamination (reads not mapping to *M. tuberculosis*) per sample.  
438 Ten samples had  $<15\%$  contaminant reads, although contamination levels

439 increased as high as 75% before the proportion of the *M. tuberculosis* genome  
440 recovered started to drop. Low numbers of *M. tuberculosis* reads could also  
441 reflect poor DNA quality from samples stored for long periods, as most of the  
442 samples with <80% reference genome coverage (12/17, 71%) were more than 3  
443 weeks old before extraction.

444

445

#### 446 **Concordance of results from direct and MGIT samples**

447

448 We took a set of 68,695 high quality SNPs obtained from analysis of 3480  
449 samples (1), and genotyped all samples at these positions (see Methods). This  
450 allowed us to calculate a genetic distance between the 17 paired MGIT and direct  
451 samples. The median (and modal) SNP difference was 1 (Figure 4a), with one  
452 outlier pair of samples that differed by 1106 SNPs, discussed below and all other  
453 differences ≤22 SNPs.

454

455 We placed 17 paired direct and MGIT samples on the phylogeny from (1) (see  
456 Methods). Our samples were distributed across global diversity (Figure 4b; tree  
457 thinned to aid visibility). For the pair with 1106 SNP differences, the MGIT  
458 sample was placed very closely to 3 other pairs (0 SNP difference to one sample,  
459 and 5 SNP differences to the others). Although this might result from different  
460 strains being present within the host, a within-laboratory labelling error or  
461 cross-contamination is also possible.

462

463

#### 464 **No evidence of higher diversity in direct samples**

465

466 Comparing direct/MGIT pairs where both samples had at least 20x mean depth  
467 of coverage on the *M. tuberculosis* reference, the median number of high  
468 confidence (see Methods) heterozygous sites was 25 in both direct and MGIT  
469 samples. There was no clear trend of greater genome-wide diversity in direct  
470 samples (Supplementary Figure 2).

471

#### 472 **Detection of *M. tuberculosis* in culture positive/negative samples**

473

474 All sequenced culture-positive (37/39) direct *M. tuberculosis* samples were  
475 successfully identified by Mykrobe predictor to complex level (37/37) and 95%  
476 to species level (35/37), including 13/37 (35%) where the mean depth of  
477 coverage was <3. All MGIT cultures were identified as *M. tuberculosis*. We were  
478 also able to identify *M. tuberculosis* in 2/2 direct samples with low AFB scores  
479 (+1) and no growth in MGIT culture; this may represent dead bacilli from a  
480 patient undergoing treatment.

481

482

#### 483 **Antibiotic resistance prediction**

484

485

486 In total 168 predictions for first-line (n=96) and second-line (n=72) antibiotic  
resistance were made for the 24/37 (65%) direct samples which had at least 3x  
depth (Supplementary Tables 4,5). For the 13/37 (35%) samples that had <3x

487 depth, no resistance predictions were made. This included 1/2 culture-negative  
488 samples.

489  
490 92/96 (96%) predictions for the first-line antibiotics were concordant with  
491 reference laboratory DST. The four mismatches (three pyrazinamide mixed  
492 genotypes with both R and S alleles present, and one rifampicin resistant  
493 genotype with sensitive phenotype) were found across three samples, all from  
494 the same patient (patient 2 in Supplementary Table 5) who had a variable  
495 phenotype for rifampicin and pyrazinamide. The resistant genotype for  
496 rifampicin was consistent across all three samples from this patient  
497 (*rpoB\_I491F*). There is evidence that this mutation causes resistance, but that the  
498 phenotype is often reported as sensitive (27,28,1). The mixed genotype for  
499 pyrazinamide was again consistent with presence of both R and S alleles on  
500 *pncA\_V7L* across all three samples, whereas the phenotype varied. This mutation  
501 is also known to confer resistance in samples reported as phenotypically  
502 sensitive (1). Further, 1/3 samples from this patient (sample 602112,  
503 Supplementary Table 4) contained two additional mutations conferring  
504 resistance to isoniazid and pyrazinamide, *katG\_S315T* and *pncA\_T135P*  
505 respectively, which were not detected in the previous or following sample. This  
506 variation between same-patient samples taken over time may represent ongoing  
507 evolution, changing population size, and within-patient diversity of *M.*  
508 *tuberculosis* as previously demonstrated by Eldholm *et al* (29). In addition to the  
509 above, WGS provided 72 predictions for second-line antibiotics where DST was  
510 not attempted.

511  
512 The 13/37 samples that yielded insufficient WGS data for resistance prediction  
513 had a higher proportion of other bacterial DNA (Figure 3b; median 96%, IQR 38-  
514 70%, vs median 12%, IQR 0-67% , in those where resistance prediction was  
515 possible, rank-sum p=0.01).

516  
517 **Sub-24 hour turnaround time with Illumina MiniSeq**  
518

519 Illumina MiniSeq sequencing for three samples (single run; 1 pure BCG, 2  
520 negative sputum DNA spiked with BCG DNA) was completed in 6 hours 40  
521 minutes. BCG reference genome coverage was 31-33x in spiked samples, and 84x  
522 in pure BCG (Table 1). In all cases the species/strain was correctly identified as  
523 *M. bovis* strain BCG, and pyrazinamide resistance was correctly identified due to  
524 mutation H57D in *pncA*.

525  
526  
527 **Modified method for ONT MinION**  
528

529 A new PCR-based rapid 1D MinION protocol was tested using extracted BCG  
530 DNA, ZN-negative sputum DNA spiked with BCG DNA, and R9 flowcells (see  
531 Methods). Analysis of genome-wide coverage distribution confirmed that use of  
532 PCR had not led to significant coverage bias (Supplementary Figure 3), and that  
533 >95% of the reference genome attained coverage >5x for all samples. In all cases  
534 Mykrobe correctly identified the species/strain as *M. bovis* strain BCG (Table 2).  
535 Amplification with Phusion High-Fidelity master mix resulted in the highest yield

536 (760Mb, with 68x coverage of BCG). The pure BCG and both 15% spike  
537 experiments resulted in correct identification of the H57D mutation in *pncA* that  
538 confers pyrazinamide resistance in BCG.

539  
540 No susceptibility predictions were made in the 5% and 10% spike experiments,  
541 owing to low coverage. The 19x and 10x coverage depth, respectively in these  
542 samples, corresponded to kmer coverages ( $k=15$ ) <3x due to sequencing errors.  
543 Higher coverage depth did permit resistance prediction; for example, the  
544 LongAmp/15% sample had 35x coverage depth, corresponding to kmer  
545 coverage of 3x at the H57D allele. For each of the 3 samples with enough  
546 coverage to allow susceptibility predictions, 200 resistance SNPs/indels were  
547 genotyped by Mykrobe and no false resistance calls were made. Using two  
548 independent methods (see Methods) we found a strong bias in the distribution of  
549 SNP errors in the consensus of the MinION data. Both methods agreed the bias  
550 was systematic though not on the precise proportions (e.g. of SNP errors in de  
551 novo assembly, 50% were A->G, and 44% were T->C; Supplementary Table 1,2),  
552 consistent with a strong A->G error bias within a 1D read. A filter to ensure SNP  
553 calls have support from reads mapping to both strands could remove such  
554 errors.

555  
556 In all 5 samples sequenced on R9 flowcells, data yield was highest at the start of  
557 the run, with consistent yield profiles. For the Phusion/15% run we obtained  
558 over 65% of the data in 8 hours, and 80% in 10 hours (Supplementary Figure 4).  
559 Thus we were able to detect BCG and predict the correct resistance with 10  
560 hours of sequencing despite the high error rate (Supplementary Figure 5).

561

562

### 563 Sub-12 hour turnaround time with ONT R9.4 MinION

564

565 We sequenced a single sample (15% BCG spiked ZN-negative sputum) on the  
566 latest R9.4 MinION flowcell (see Methods). Yield was 1.3Gb in 48 hours; we were  
567 able to detect *M. tuberculosis* complex and identify the strain as BCG within 1  
568 hour of sequencing, and correctly predict pyrazinamide resistance after 2 further  
569 hours. The additional yield would enable 3-5 samples to be sequenced per flow  
570 cell. Currently, multiplexing capabilities are limited by contamination, and the  
571 kmer-coverage attained on susceptible/resistance alleles.

572

573 Although in this instance we performed base-calling (conversion of raw MinION  
574 electrical output to DNA reads) after sequencing completed (taking 48 hours on  
575 a quad-core desktop computer), we subsequently verified (on other samples)  
576 that real-time base-calling can be performed using ONTs MinKNOW software.  
577 This would give a turnaround time of 8 hours (Figure 5).

578

579 We took our phylogenetic placement of the MiniSeq BCG data as truth, 4 SNPs  
580 distant from a BCG sample on the predefined tree. By comparison, after 1 hour of  
581 sequencing with R9.4, we were able to genotype 15592 of the 68695 SNPs,  
582 placing the sample at the correct leaf of the tree, at an estimated distance of 25  
583 SNPs. Thus, our genotyping on 1D nanopore reads had at most 29 errors out of  
584 15592 SNPs - an error rate of 0.2%.

585

## 586 Costing

587 Reagent costs (sample decontamination, extraction, sequencing library  
588 preparation, and sequencing) per sample were £96 (MiSeq, 12 samples/run),  
589 £198 (MiniSeq, 3 samples/run), £515 (R9 MinION, 1 sample/run), £101-172  
590 (R9.4 MinION, between 3 and 5 samples/run, approximate cost as multiplexing  
591 kit not yet available). See Supplementary Table 3 for details.

592

593

## 594 Discussion

595

596 Anticipating a growing knowledge-base of the molecular determinants of  
597 antibiotic resistance (1), we have developed a method of extracting and purifying  
598 mycobacterial DNA from primary clinical samples and producing accurate  
599 sequence data in clinically a useful timeframe. We have demonstrated: first, that  
600 direct WGS of sputum is possible, and gives concordant DST results with  
601 phenotype and concordant phylogenetic placement with culture-based  
602 sequencing. Second: using an Illumina MiSeq sequencer, we can obtain results  
603 within 48 hours for <£100 per sample. Using an Illumina MiniSeq can deliver a  
604 same-day test (16 hours) for an estimated consumable cost of £198 per sample.  
605 Although the costs presented here only represent reagents, they are well below  
606 that of traditional phenotyping (£545 to provide first-/second- line DST and  
607 MIRU-VNTR in a bottom-up costing including, for example, consumables, staff  
608 time, and overheads (16)). The MiSeq cost is also well below that of the  
609 SureSelect procedure (£203 per sample) (18).

610

611 The World Health Organisation (WHO) has called for affordable and accessible  
612 point-of-care TB diagnostics, including for DST. Current molecular assays  
613 provide partial information on some drugs, but do not easily scale to incorporate  
614 a growing list of recognized resistance mutations. Furthermore, additional  
615 assays are currently needed where surveillance or outbreak detection are  
616 indicated, at additional cost. A single assay providing diagnostic information, and  
617 data for surveillance and outbreak detection is therefore an attractive prospect.

618

619 In cities where there are large numbers of TB cases (for example upward of  
620 65,000 TB cases per years in Mumbai) centralized sequencing services taking  
621 advantage of high throughput Illumina sequencing platforms may be applicable.  
622 However, the relatively high capital costs, and requirement for a well-equipped  
623 laboratory are an impediment to implementation across the full range of  
624 locations across the world. For a complete solution, the ability to function in  
625 varied low-tech environments is a practical necessity. The MinION delivers this,  
626 as demonstrated in Guinea last year during the Ebola outbreak (30). We confirm  
627 here that, despite the high error rate in reads, given deep coverage, it is possible  
628 to accurately genotype resistance SNPs and indels using the MinION method  
629 applied here.

630

631 Since with Illumina technology the depth of sequencing is determined in  
632 advance, the small amount of *M. tuberculosis* in a direct sample can result in test  
633 failures. In this experiment *M. tuberculosis* identification and susceptibility

634 prediction failed in 2/39 and 13/37 samples respectively. MinION sequencing  
635 allows sequencing to continue until sufficient coverage has been obtained, giving  
636 faster results when there is high load, and avoiding this type of failure when the  
637 load is low. The throughput obtained here with R9.4 flowcells (1.3 Gb), would  
638 allow an 8 hour turnaround time with one sample per flowcell. Overall yield gave  
639 BCG coverage depth of 147x, sufficient to have multiplexed 3-5 samples, with  
640 implied cost of around £101-172/sample – more data is needed to confidently  
641 determine a realistic multiplexing level. Although these estimated costs are not  
642 yet at a level to be affordable in routine global use, there is clearly scope for  
643 further technology-driven cost reduction, either through improved  
644 enrichment/depletion (Figure 2a), higher sequencing yield and multiplexing, or  
645 real-time filtering of contamination (31).

646  
647 Direct from patient sequencing may also allow identification of mycobacterial co-  
648 colonization, which may occur in 3-5% of patients (32). Indeed, since infections  
649 are chronic and structured in the lung, it is expected that transmission models  
650 will need to account for mixtures and within-host evolution (33).

651  
652 Diagnostic and surveillance information direct from patient specimens can now  
653 be obtained in as little as 8 hours on the ONT sequencing platform, and in 16/44  
654 hours on Illumina platforms, a considerable step forward. Faster and more  
655 automated sample processing, as well as a cost reduction, is a clear necessity for  
656 global take-up in low-income settings. Achieving this would revolutionise the  
657 management of TB.

658

659

## 660 Acknowledgements

661 We thank Phuong Quan for assistance with statistical analysis, Rachel Norris for  
662 help with error analysis, Kevin Hall and Aurelie Modat from Illumina for helping  
663 with the MiniSeq sequencing, and David Stoddart and Oliver Hartwell from  
664 Oxford Nanopore Technologies for giving us help and early access to the rapid  
665 PCR 1D prep.

666

667

668

## 669 References

- 670  
671 1. Walker TM, Kohl TA, Omar SV, Hedge J, del Ojo Elias C, Bradley P, Iqbal Z,  
672 Feuerriegel S, Niehaus KE, Wilson DJ, Clifton DA, Kapatai G, Ip CLC, Bowden  
673 R, Drobniowski FA, Allix-Beguec C, Gaudin C, Parkhill J, Diel R, Supply P,  
674 Crook DW, Smith EG, Walker AS, Ismail N, Niemann S, Peto TEA,  
675 **Modernising Medical Microbiology (MMM) Informatics Group** (2015)  
676 Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug  
677 susceptibility and resistance: a retrospective cohort study. *The Lancet Infectious*  
678 *Diseases* 15:1193-202  
679  
680 2. Said HM, Kock MM, Ismail NA, Baba K, Omar SV, Osman AG, Hoosen AA,  
681 Ehlers MM (2012) Evaluation of the GenoType MTBDRsl assay for susceptibility

- 682 testing of second-line anti-tuberculosis drugs. Int J Tuberc Lung Dis. 2012  
683 Jan;16(1):104-9
- 684
- 685 **3. WHO Expert Group Report** (2008) Molecular Line Probe Assays for Rapid  
686 Screening of Patients at Risk of Multi-Drug Resistant Tuberculosis (MDR-TB)
- 687
- 688 **4. Colman RE, Anderson J, Lemmer D, Lehmkuhl E, Georghiou SB, Heaton H,**  
689 **Wiggins K, Gillece JD, Schupp JM, Catanzaro DG, Crudu V, Cohen T, Rodwell**  
690 **TC, Engelthaler DM** (2016) Rapid drug susceptibility testing of drug-resistant  
691 *Mycobacterium tuberculosis* isolates directly from clinical samples by use of  
692 amplicon sequencing: a concept study. J Clin Microbiol 54:2058-2067
- 693
- 694 **5. Lee RS, Behr MA** (2016) The implications of whole-genome sequencing in the  
695 control of tuberculosis. Therapeutic advances in infectious disease 3:47-62
- 696
- 697 **6. Takiff HE, Feo O** (2015) Clinical value of whole-genome sequencing of  
698 *Mycobacterium tuberculosis*. Lancet Infect Dis 15:1077-90
- 699
- 700 **7. Witney AA, Cosgrove CA, Arnold A, Hinds J, Stoker NG, Butcher PD** (2016)  
701 Clinical use of whole genome sequencing for *Mycobacterium tuberculosis*. BMC  
702 medicine 14:46
- 703
- 704 **8. Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O,**  
705 **Konsevaya I, Corander J, Bryant J, Parkhill J, Nejentsev S, Horstmann RD,**  
706 **Brown T, Drobniowski F** (2014) Evolution and transmission of drug-resistant  
707 tuberculosis in a Russian population. Nat Genet 46:279-86
- 708
- 709 **9. Clark TG, Mallard K, Coll F, Preston M, Assefa S, Harris D, Ogwang S,**  
710 **Mumbowa F, Kirenga B, O'Sullivan DM, Okwera A, Eisenach KD, Joloba M,**  
711 **Bentley SD, Ellner JJ, Parkhill J, Jones-Lopez EC, McNerney R** (2013)  
712 Elucidating emergence and transmission of multidrug-resistant tuberculosis in  
713 treatment experienced patients by whole genome sequencing. PLoS One  
714 8:e83012
- 715
- 716 **10. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC,**  
717 **Warren RM, Streicher EM, Calver A, Sloutsky A, Kaur D, Posey JE, Plikaytis**  
718 **B, Oggioni MR, Gardy JL, Johnston JC, Rodrigues M, Tang PKC, Kato-Maeda**  
719 **M, Borowsky ML, Muddukrishna B, Kreiswirth BN, Kurepina N, Galagan J,**  
720 **Gagneux S, Birren B, Rubin EJ, Lander ES, Sabeti PC, Murray M** (2013)  
721 Genomic analysis identifies targets of convergent positive selection in drug-  
722 resistant *Mycobacterium tuberculosis*. Nat Genet 45:1183-9
- 723
- 724 **11. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S,**  
725 **Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K,**  
726 **Jones SJM, Brinkman FSL, Brunham RC, Tang P** (2011) Whole-genome  
727 sequencing and social-network analysis of a tuberculosis outbreak. The New  
728 England journal of medicine 364:730-9
- 729

- 730    12. **Guerra-Assuncao JA, Crampin AC, Houben RM, Mzembe T, Mallard K,**  
731    **Coll F, Khan P, Banda L, Chiwaya A, Pereira RP, McNerney R, Fine PE,**  
732    **Parkhill J, Clark TG, Glynn JR** (2015) Large-scale whole genome sequencing of  
733    *M. tuberculosis* provides insights into transmission in a high prevalence area.  
734    *Elife*. 2015 Mar 3;4.
- 735
- 736    13. **Stucki D, Ballif M, Bodmer T, Coscolla M, Maurer AM, Droz S, Butz C,**  
737    **Borrell S, Langle C, Feldmann J, Furrer H, Mordasini C, Helbling P, Rieder**  
738    **HL, Egger M, Gagneux S, Fenner L** (2015) Tracking a tuberculosis outbreak  
739    over 21 years: strain-specific single-nucleotide polymorphism typing combined  
740    with targeted whole-genome sequencing. *J Infect Dis* 211:1306-16
- 741
- 742    14. **Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre**  
743    **DW, Wilson DW, Hawkey PM, Crook DW, Parkhill J, Harris D, Walker AS,**  
744    **Bowden R, Monk P, Smith EG, Peto TE** (2013) Whole-genome sequencing to  
745    delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational  
746    study. *Lancet Infect Dis* 13:137-46
- 747
- 748    15. **Walker TM, Lalor MK, Broda A, Saldana Ortega L, Morgan M, Parker L,**  
749    **Churchill S, Bennett K, Golubchik T, Giess AP, Del Ojo Elias C, Jeffery KJ,**  
750    **Bowler IC, Laurenson IF, Barrett A, Drobniowski F, McCarthy ND, Anderson**  
751    **LF, Abubakar I, Thomas HL, Monk P, Smith EG, Walker AS, Crook DW, Peto**  
752    **TE, Conlon CP** (2014) Assessment of *Mycobacterium tuberculosis* transmission  
753    in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an  
754    observational study. *Lancet Respir Med* 2014 Apr;2(4):285-92
- 755
- 756    16. **Pankhurst LJ, Del Ojo Elias C, Votintseva AA, Walker TM, Cole K, Davies J,**  
757    **Fermont JM, Gascoyne-Binzi DM, Kohl TA, Kong C, Lemaitre N, Niemann S,**  
758    **Paul J, Rogers TR, Roycroft E, Smith EG, Supply P, Tang P, Wilcox MH,**  
759    **Wordsworth S, Wyllie D, Xu L, Crook DW, COMPASS-TB Study Group** (2016)  
760    Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-  
761    genome sequencing: a prospective study. *Lancet Respir Med* 2016 Jan;4(1):49-58
- 762
- 763    17. **Doughty EL, Sergeant MJ, Adetifa I, Antonio M, Pallen MJ** (2014) Culture-  
764    independent detection and characterisation of *Mycobacterium tuberculosis* and  
765    *M. africanum* in sputum samples using shotgun metagenomics on a benchtop  
766    sequencer. *PeerJ* 2014 Sep 23;2:e585
- 767
- 768    18. **Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZ,**  
769    **Depledge DP, Nikolayevskyy V, Broda A, Stone MJ, Christiansen MT,**  
770    **Williams R, McAndrew MB, Tutill H, Brown J, Melzer M, Rosmarin C,**  
771    **McHugh TD, Shorten RJ, Drobniowski F, Speight G, Breuer J** (2015) Rapid  
772    Whole-Genome Sequencing of *Mycobacterium tuberculosis* Isolates Directly from  
773    Clinical Samples. *J Clin Microbiol* 53:2230-7
- 774
- 775    19. **Votintseva AA, Pankhurst LJ, Anson LW, Morgan MR, Gascoyne-Binzi D,**  
776    **Walker TM, Quan TP, Wyllie DH, Del Ojo Elias C, Wilcox M, Walker AS, Peto**  
777    **TE, Crook DW** (2015) Mycobacterial DNA extraction for whole-genome

- 778 sequencing from early positive liquid (MGIT) cultures. J Clin Microbiol 53:1137-  
779 43  
780  
781 20. **Li, H.** (2013) Aligning sequence reads, clone sequences and assembly contigs  
782 with BWA- MEM. arXiv:1303.3997  
783  
784 21. **Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, Earle S,**  
785 **Pankhurst LJ, Anson L, de Cesare M, Piazza P, Votintseva AA, Golubchik T,**  
786 **Wilson DJ, Wyllie DH, Diel R, Niemann S, Feuerriegel S, Kohl TA, Ismail N,**  
787 **Omar SV, Smith EG, Buck D, McVean G, Walker AS, Peto TE, Crook DW, Iqbal**  
788 **Z** (2015) Rapid antibiotic-resistance predictions from genome sequence data for  
789 *Staphylococcus aureus* and *Mycobacterium tuberculosis*. Nat. Commun. 2015 Dec  
790 21;6:10063  
791  
792 22. **Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G** (2012) De novo  
793 assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet  
794 2012 Jan 8; 44(2):226-232  
795  
796 23. **Loman NJ, Quinlan AR** (2014) Poretools: a toolkit for analysing nanopore  
797 sequence data. Bioinformatics 2014 Dec 1;30(23):3399-401  
798  
799 24. **Vaser R, Sovic I, Nagarajan N, Sikic M** (2016) Fast and accurate de novo  
800 assembly from long uncorrected reads bioRxiv  
801 <http://dx.doi.org/10.1101/068122>  
802  
803 25. **Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C,**  
804 **Salzberg S** (2004) Versatile and open software for comparing genomes  
805  
806 26. **Koren S, Walenz BP, Berlin K, Miller JR, Phillippy AM** (2016) Canu:  
807 scalable and accurate long read assembly via adaptive k-mer weighting and  
808 repeat separation. bioRxiv: <http://dx.doi.org/10.1101/071282>  
809  
810 27. **Cohen KA, Abeel T, Manson McGuire A, et al.** (2015) Evolution of  
811 Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome  
812 Sequencing and Dating Analysis of *Mycobacterium tuberculosis* Isolates from  
813 KwaZulu-Natal. PLoS medicine 12:e1001880  
814  
815 28. **Sanchez-Padilla E, Merker M, Beckert P, et al.** (2015) Detection of drug-  
816 resistant tuberculosis by Xpert MTB/RIF in Swaziland. N Engl J Med 2015 Mar  
817 19;372(12):1181-2  
818  
819 29. **Eldholm V, Norheim G, Lippe Bvd, Kinander W, Dahle UR, Caugant DA,**  
820 **Mannsaker T, Mengshoel AT, Dyrhol-Riise AM, Balloux F** (2014) Evolution of  
821 extensively drug-resistant *Mycobacterium tuberculosis* from a susceptible  
822 ancestor in a single patient. Genome Biol. 2014; 15(11): 490.  
823  
824 30. **Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA,**  
825 **Koundouno R, Dudas G, Mikhail A, Ouédraogo N, Afrough B, Bah A, Baum JH,**  
826 **Becker-Ziaja B, Boettcher JP, Cabeza-Cabrerozo M, Camino-Sánchez Á,**

827 **Carter LL, Doerrbecker J, Enkirch T, García-Dorival I, Hetzelt N, Hinzmann J,**  
828 **Holm T, Kafetzopoulou LE, Koropogui M, Kosgey A, Kuisma E, Logue CH,**  
829 **Mazzarelli A, Meisel S, Mertens M, Michel J, Ngabo D, Nitzsche K, Pallasch E,**  
830 **Patrono LV, Portmann J, Repits JG, Rickett NY, Sachse A, Singethan K,**  
831 **Vitoriano I, Yemanaberhan RL, Zekeng EG, Racine T, Bello A, Sall AA, Faye O,**  
832 **Faye O, Magassouba N, Williams CV, Amburgey V, Winona L, Davis E,**  
833 **Gerlach J, Washington F, Monteil V, Jourdain M, Bererd M, Camara A,**  
834 **Somlare H, Camara A, Gerard M, Bado G, Baillet B, Delaune D, Nebie KY,**  
835 **Diarra A, Savane Y, Pallawo RB, Gutierrez GJ, Milhano N, Roger I, Williams**  
836 **CJ, Yattara F, Lewandowski K, Taylor J, Rachwal P, Turner DJ, Pollakis G,**  
837 **Hiscox JA, Matthews DA, O'Shea MK, Johnston AM, Wilson D, Hutley E, Smit**  
838 **E, Di Caro A, Wölfel R, Stoecker K, Fleischmann E, Gabriel M, Weller SA,**  
839 **Koivogui L, Diallo B, Keïta S, Rambaut A, Formenty P, Günther S, Carroll MW**  
840 **(2016) Real-time, portable genome sequencing for Ebola surveillance. Nature**  
841 **2016 Feb 11;530(7589):228-32**

842

843 **31. Loose M, Malla S, Stout M (2016): Real-time selective sequencing using**  
844 **nanopore technology. Nat Methods 2016 Sep;13(9):751-4**

845

846 **32. Wyllie D, Bawa Z, Walker T, Peto T, Smoth G (2016): Low frequency of**  
847 **mixed *M. tuberculosis* infection in a large prospective UK series as assessed using**  
848 **whole genome sequencing. 26<sup>th</sup> ECCMID Amsterdam, Netherlands, 9-12 April**  
849 **2016 P0153**

850

851 **33. Lieberman T, Wilson D, Misra R, Xiong LL, Moodley P, Cohen T, Kishony**  
852 **R (2016)**

853 Genomic diversity in autopsy samples reveals within-host dissemination of HIV-  
854 associated Mycobacterium tuberculosis. Nat Med 2016 Oct 31

855

856

857

858

859

860

861

862

863

864

865

866

867

868

869

870

871

872

873

874

875

876

877 **Table 1: Yield from pure BCG, and from negative sputum spiked with BCG -**  
878 **sequenced on Illumina MiniSeq**

879

880

	estimated Fmol loaded	Yield (Mb)	Read length (bp)	BCG reference coverage
Pure BCG TB1_N716	800	381	101	84.0
50 % BCG TB1_N718	800	244	101	31.0
50% BCG TB1_N719	800	257	101	33.0

881

882

883

884

885 **Table 2: Yield from pure BCG, and from negative sputum spiked with BCG -**  
886 **both sequenced with MinION 1D protocol**

887

888

Model	Sample	Fmol loaded	Read count	Yield /Mb	Avg read length/kb	BCG covg depth	SNPs (kmer covg on R allele)
R9	Pure culture d BCG	Not available	297,239	360	1.2	80	H57D (13)
R9	5% BCG LongA mp	82	182,670	559	2.0	19	No calls
R9	10% BCG LongA mp	76	180,507	467	1.8	10	No calls
R9	15% BCG LongA mp	51	203,285	627	2.0	35	H57D (3)
R9	15% BCG Phusion	27	184,895	758	2.4	68	H57D (8)
R9.4	15% BCG Phusion	43	754,338	1306	1.7	147	H57D (20)

889

890

891 **Figures**

892

Figure 1

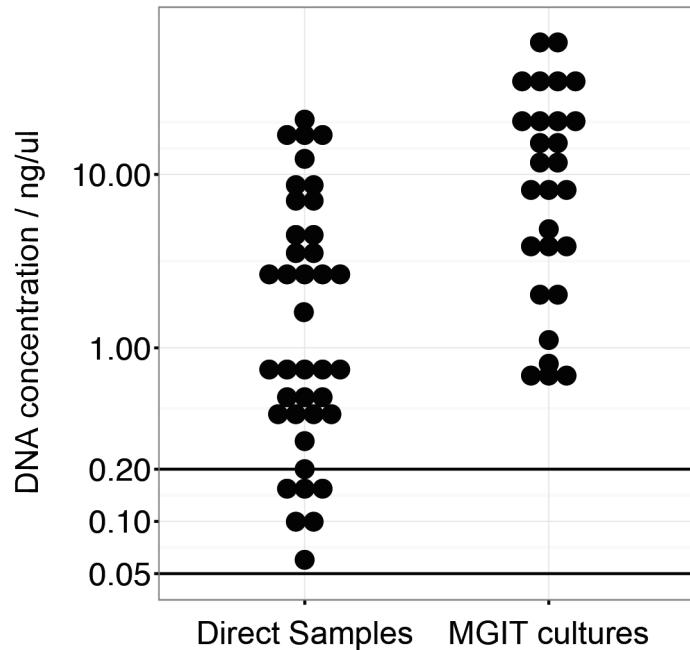


Figure 1: DNA extracted (ng/μl) from MGIT cultures and direct clinical samples. Each dot represents a single extraction. Horizontal line at 0.2 ng/μl represents the DNA concentration required for MiSeq library preparation. Horizontal line at 0.05 ng/μl represents minimum DNA concentration used for MiSeq library preparation from direct samples.

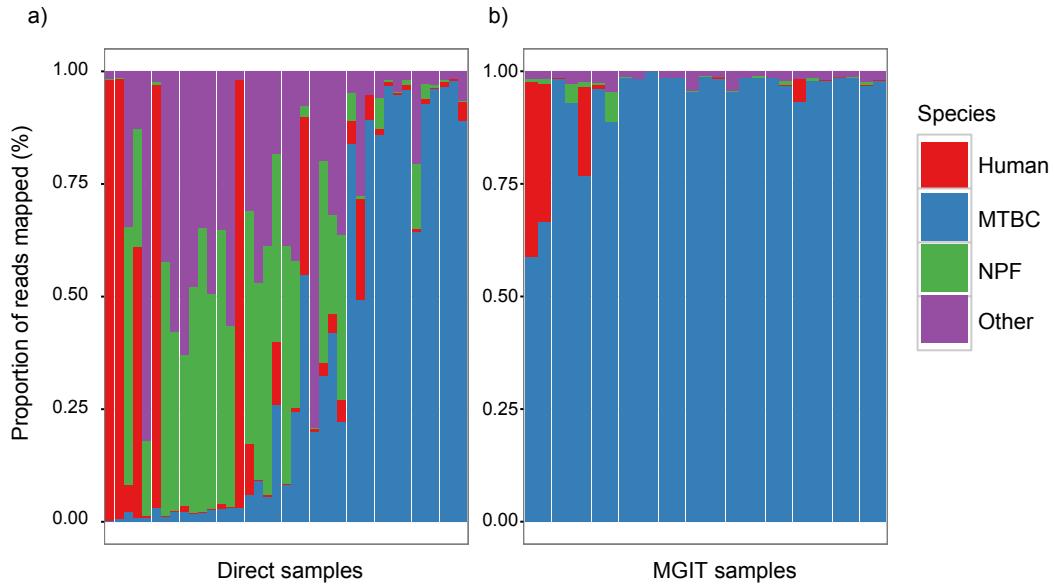
893

894

895

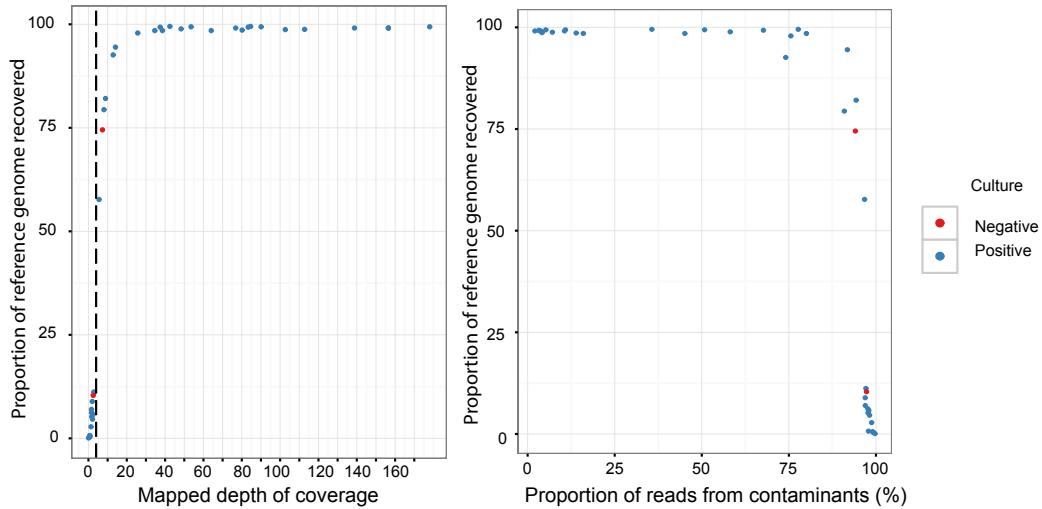
896

897 **Figure 1:** DNA extracted (ng/μl) from MGIT cultures and direct clinical samples.  
898 Each dot represents a single extraction. Horizontal line at 0.2 ng/μl represents  
899 the DNA concentration required for MiSeq library preparation. Horizontal line at  
900 0.05 ng/μl represents minimum DNA concentration used for MiSeq library  
901 preparation from direct samples.



902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
920  
921  
922  
923  
924

**Figure 2:** Proportion of reads assigned to various species categories in each sample. **a)** Direct samples show removal of human DNA (red) has been broadly successful, but removal of naso-pharyngeal flora (NPF, green) and other bacteria (purple) had more variable success. **b)** MGIT samples show much more uniform dominance of *M. tuberculosis* reads, as expected after 2 weeks of culture designed to favour mycobacterial growth.

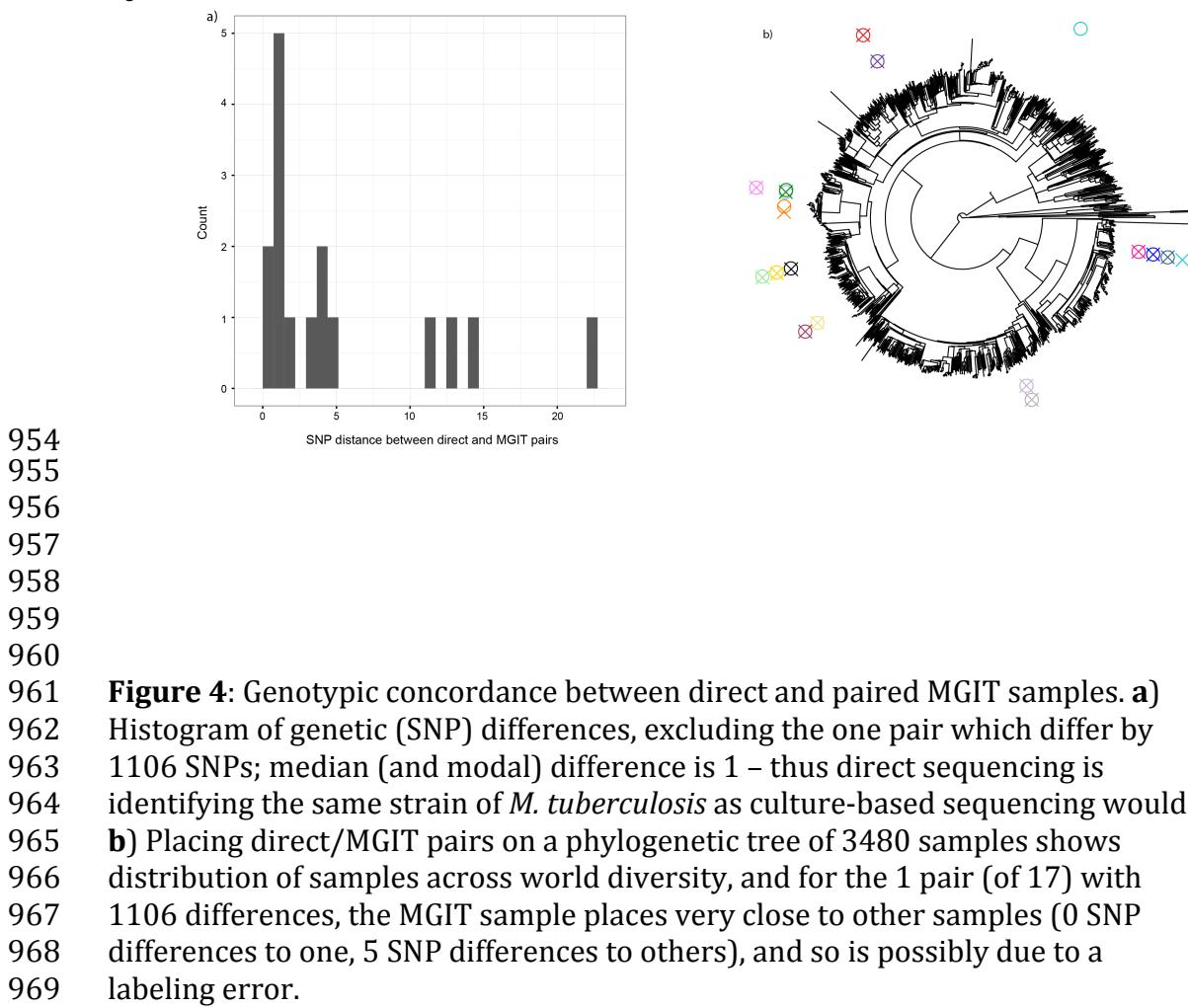


925  
926  
927  
928  
929  
930

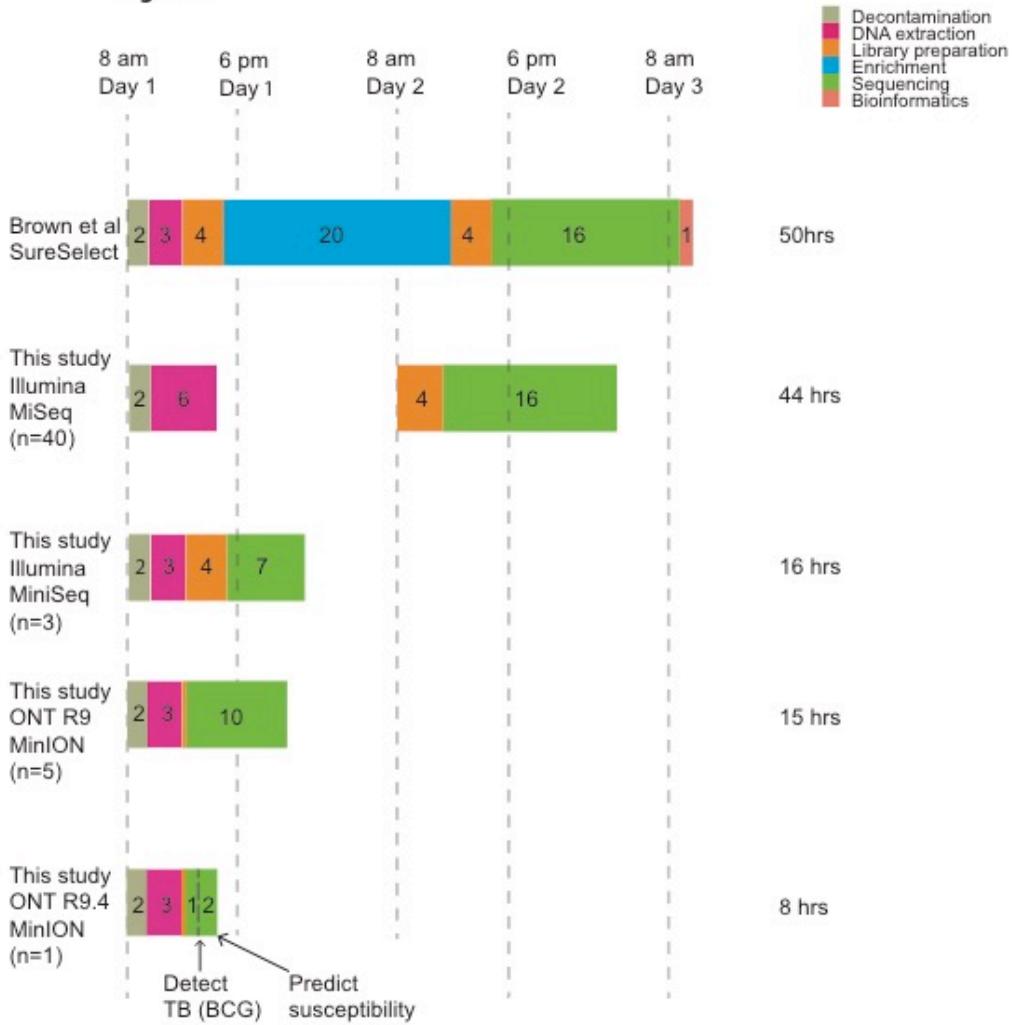
931 **Figure 3:** Recovery of *M. tuberculosis* genome in direct samples and robustness  
932 to contamination. **a)** Depth versus proportion of the *M. tuberculosis* reference  
933 recovered (at >5x depth). Samples either have more than 10x depth and recover  
934 more than 90% of the genome, or have <3x depth and recover less than 12% of  
935 the genome. Vertical dotted line at 3x depth is threshold for susceptibility  
936 prediction. **b)** Proportion of contamination (reads not mapping to *M. tuberculosis*  
937 reference) versus proportion of genome recovered. Samples with less than 95%  
938 of the *M. tuberculosis* genome recovered all have >75% of contamination.

939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953

Figure 4



**Figure 5**



970

971

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

**Figure 5:** Timelines and cost. We compare the method of Brown et al with the results of this study, using the Illumina MiSeq and MiniSeq, and the ONT MinION. We assume that no step of the process can be initiated after 6pm or before 8am. The method of Brown et al has a rapid extraction step, but also a 20 hour enrichment step, resulting in a 50 hour turn-around time, at an estimated cost of £203/sample. By contrast, our extraction method with MiSeq sequencing provides results at £94/sample. The DNA extraction process was updated for the MiniSeq and MinION experiments, removing the ethanol precipitation step. In normal use this would take 3 hours. The thin orange rectangle on the MinION timelines is the 10 minute sample preparation step. In this experiment, since we used spiked BCG DNA in sputum, we did not use a human depletion step, thus taking only 2 hours. This figure is intended to show comparable real-use timelines, and so the MiniSeq/MinION timelines are shown with 3 hour extraction steps. The MiniSeq enables a 16-hour turnaround time, by sequencing for only 7 hours. The R9 MinION also delivers sub-24 hour results, but requires one flow-cell per sample. The R9.4 MinION gives an 8 hour turnaround time (3 hours of sequencing with real-time (i.e. simultaneous) basecalling when used on a single sample). We do not display the 48 hours we actually took to basecall the data after the run, as this was in effect our error - we subsequently confirmed (on other samples) that real-time basecalling would have worked.

991  
992  
993  
994