

1 **Title:**

2 Signatures of transmission in within-host *M. tuberculosis* variation

3

4 **Authors**

5 Katharine S. Walter¹, Ted Cohen², Barun Mathema³, Caroline Colijn⁴, Benjamin Sobkowiak², Iñaki

6 Comas⁵, Galo A. Goig^{6,7}, Julio Croda^{2,8,9}, Jason R. Andrews¹⁰

7

8 1. Division of Epidemiology, University of Utah, Salt Lake City, UT 84105

9 2. Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven,

10 USA

11 3. Department of Epidemiology, Columbia University Mailman School of Public Health; New

12 York, United States.

13 4. Department of Mathematics, Simon Fraser University; Burnaby, Canada.

14 5. Institute of Biomedicine of Valencia (CSIC), Valencia, Spain

15 6. Swiss Tropical and Public Health Institute, Allschwil, Switzerland.

16 7. University of Basel, Basel, Switzerland.

17 8. Federal University of Mato Grosso do Sul - UFMS, Campo Grande, MS, Brazil

18 9. Oswaldo Cruz Foundation Mato Grosso do Sul, Mato Grosso do Sul, Brazil

19 10. Division of Infectious Diseases and Geographic Medicine, Stanford University School of

20 Medicine, Stanford, CA, USA

21

22

23

24 **Abstract**

25 **Background.**

26 Because *M. tuberculosis* evolves slowly, transmission clusters often contain multiple individuals with
27 identical consensus genomes, making it difficult to reconstruct transmission chains. Finding additional
28 sources of shared *M. tuberculosis* variation could help overcome this problem. Previous studies have
29 reported *M. tuberculosis* diversity within infected individuals; however, whether within-host variation
30 improves transmission inferences remains unclear.

31

32 **Methods.**

33 To evaluate the transmission information present in within-host *M. tuberculosis* variation, we re-analyzed
34 publicly available sequence data from three household transmission studies, using household membership
35 as a proxy for transmission linkage between donor-recipient pairs.

36

37 **Findings.**

38 We found moderate levels of minority variation present in *M. tuberculosis* sequence data from cultured
39 isolates that varied significantly across studies (mean: 6, 7, and 170 minority variants above a 1% minor
40 allele frequency threshold, outside of PE/PPE genes). Isolates from household members shared more
41 minority variants than did isolates from unlinked individuals in the three studies (mean 98 shared
42 minority variants vs. 10; 0.8 vs. 0.2, and 0.7 vs. 0.2, respectively). Shared within-host variation was
43 significantly associated with household membership (OR: 1.51 [1.30,1.71], for one standard deviation
44 increase in shared minority variants). Models that included shared within-host variation improved the
45 accuracy of predicting household membership in all three studies as compared to models without within-
46 host variation (AUC: 0.95 *versus* 0.92, 0.99 *versus* 0.95, and 0.93 *versus* 0.91).

47

48 **Interpretation.**

49 Within-host *M. tuberculosis* variation persists through culture and could enhance the resolution of
50 transmission inferences. The substantial differences in minority variation recovered across studies
51 highlights the need to optimize approaches to recover and incorporate within-host variation into
52 automated phylogenetic and transmission inference.

53

54 **Funding**

55 NIAID: 5K01AI173385

56

57 **Keywords**

58 *M. tuberculosis*, genomics, epidemiology, transmission

59 **Introduction**

60

61 Reducing the global burden of tuberculosis urgently requires reducing the number of incident *M.*
62 *tuberculosis* infections. Yet the long and variable latency period of TB infection makes it challenging to
63 identify sources of transmission and thus intervene. Genomic epidemiology approaches have been
64 powerfully applied to characterize *M. tuberculosis* global phylogenetic structure, migration and gene
65 flow, patterns of antibiotic resistance, transmission linkages. Yet transmission inference approaches have
66 often failed to identify the majority of transmission linkages in high-incidence settings¹⁻⁶. Further, while
67 previous studies have identified heterogeneity in the number of secondary cases generated by infectious
68 individuals⁷ and risk factors for onwards transmission^{8,9}, these are often difficult to generalize. Many
69 critical questions, including the contribution of asymptomatic individuals to transmission, remain
70 unanswered. Novel, accessible approaches to reconstruct high-resolution transmission patterns are
71 urgently needed so that public health programs can identify environments driving transmission and risk
72 factors for onwards transmission.

73 Commonly used approaches for *M. tuberculosis* transmission inference use single consensus
74 genomes, representing the sequence of the most frequent alleles, from infected individuals. Closely
75 related pathogen genomes are predicted to be more closely linked in transmission chains. For example, *M.*
76 *tuberculosis* consensus sequences within a given genetic distance¹⁰⁻¹³ are considered clustered and
77 potentially epidemiologically linked. However, *M. tuberculosis* evolves at a relatively slow rate¹⁴. The
78 result is that there may be limited diversity in outbreaks. Indeed, several genomic epidemiology studies
79 reported that multiple individuals harbored identical *M. tuberculosis* genomes^{13,15-17}, making it difficult to
80 reconstruct who infected whom. This highlights a need to recover more informative variation from
81 pathogen genomes. This challenge is not unique to *M. tuberculosis*—COVID-19 outbreak investigations
82 frequently generate large numbers of identical genomes¹⁸, indicating a broad need for higher-resolution
83 pathogen genomics approaches.

84 Population-level bacterial diversity within an individual, or within-host heterogeneity, can be
85 attributed to mixed infections, infections with more than one distinct *M. tuberculosis* genotype, or *de*
86 *novo* evolution, mutations that are introduced over the course of an individual's infection¹⁹. Previous
87 research has found that a substantial proportion (10-20%)¹⁹ of infected individuals harbor mixed
88 infections with genetically diverse populations of *M. tuberculosis*¹⁹⁻²³. A portion of within-host
89 heterogeneity is likely transmitted onwards^{24,25} and therefore, within-host diversity captures potentially
90 valuable epidemiological information about transmission history^{25,26}. Complex infections are also
91 important clinically. Within-host heterogeneity is associated with poor treatment outcomes^{20,27} and hetero-
92 resistance, the presence of both resistant and susceptible alleles within a single infection, reduces the
93 accuracy of diagnostics for antibiotic resistance²⁷.

94 Despite the evidence that within-host *M. tuberculosis* variation is common, there are many open
95 questions about whether shared within-host variation is a predictor of transmission linkage and, more
96 practically, how to recover this level of variation and incorporate it into transmission inferences.
97 Currently, *M. tuberculosis* is most frequently cultured from sputum samples and sequenced with short
98 reads to generate a single consensus sequence²⁸. This limits the variation recovered because (a) culture
99 imposes a severe bottleneck²⁹⁻³¹; (b) within-host variation, including mixed infections, are often excluded,
100 in part due to a lack of validated methodological approaches for accurate recovery of such variation^{25,31};
101 and (c) repetitive genomic regions, including the PE/PPE gene families, among the most variant-rich and
102 potentially informative regions of the genome, are excluded³²⁻³⁴.

103 Recent work has demonstrated that pathogen enrichment approaches—through either host DNA
104 depletion or pathogen DNA enrichment—can allow *M. tuberculosis* to be sequenced directly from clinical
105 samples, bypassing the need for culture^{23,35-39}. But there have not been consistent findings about whether
106 culture-free approaches improve the detection of within-host variation.

107 *M. tuberculosis* transmission is never directly observed, making it difficult to assess the
108 performance of genomic methods in identifying true transmission pairs. We therefore tested whether
109 household members—as a proxy for epidemiologically linked individuals—shared more minority variants

110 than did unlinked individuals, and whether minority variation could enhance transmission inferences by
111 re-analyzing previously published household transmission studies.

112 Here, we study household members as a gold standard, the best available proxy for transmission
113 pairs, to test whether shared minority *M. tuberculosis* variation may augment fixed genomic differences in
114 reconstructing epidemiological linkages. In practice, such as in routine population-wide genomic *M.*
115 *tuberculosis* sequencing by public health laboratories, epidemiological linkages are frequently unknown.
116 Whether a signal of shared minority variation exists in gold standard transmission pairs can then indicate
117 whether shared minority variation might contribute to resolving such unobserved transmission linkages in
118 population-wide genomic data.

119

120 **Methods**

121

122 *Sequence and epidemiological data*

123 We accessed publicly available data from 3 household transmission studies for which both raw
124 sequence data and epidemiological linkages were publicly available: Colangeli et al. (2020)⁴⁰, Vitória,
125 Brazil; Guthrie et al. (2018)⁴¹, British Columbia, Canada; and Walker et al. 2014, Oxfordshire, England¹¹
126 (Table 1). Sequence data was available from the Sequence Read Archive (PRJNA475130,
127 PRJNA413593, and PRJNA549270). Colangeli et al. cultured sputa on Lowenstein-Jensen (LJ) slants,
128 plated cultures on Middlebrook 7H10 agar, and then scraped three loops of culture for DNA extraction⁴⁰.
129 Both Guthrie et. al and Walker et al. re-cultured frozen isolates on MGIT liquid medium or LJ slants^{11,41}.

130 We accessed information on household pairs from published phylogenies in the Colangeli et al. and
131 Guthrie et al. papers. For the Walker et al. paper, household linkages were available in the data
132 supplement.

133

134 *Bioinformatic analysis*

135 We processed raw sequence data with a previously described variant identification pipeline
136 available on GitHub (<https://github.com/ksw9/mtb-call2>).^{40,42} We previously conducted a variant
137 identification experiment to compare commonly used mapping and variant calling algorithms in *M.*
138 *tuberculosis* genomic epidemiology³². We found that the combination of the *bwa*⁴³ mapping algorithm
139 and *GATK*^{44,45} variant caller routinely minimizes false positive variant calls with minimal cost to
140 sensitivity as compared to other tool combinations⁴⁶, especially when the PE/PPE genes are excluded. We
141 therefore used this combination of tools in our pipeline.

142 Briefly, we trimmed low-quality bases (Phred-scaled base quality < 20) and removed adapters
143 with Trim Galore v. 0.6.5 (stringency=3)⁴⁷. We used CutAdapt v.4.2 to further filter reads (--nextseq-
144 trim=20 --minimum-length=20 --pair-filter=any)⁴⁸. To exclude potential contamination which a previous
145 study shows can be a source of false genetic variation⁴⁹, we used Kraken2 to taxonomically classify reads
146 and remove reads that were not assigned to the *Mycobacterium* genus or that were assigned to a
147 *Mycobacterium* species other than *M. tuberculosis*⁵⁰. We mapped reads with *bwa* v. 0.7.15 (*bwa mem*)⁴³
148 to the H37Rv reference genome (NCBI Accession: NC_000962.3
149 [https://www.ncbi.nlm.nih.gov/nuccore/NC_000962.3]) and removed duplicates with sambamba⁵¹. We
150 called variants with GATK 4.1 HaplotypeCaller⁴⁴, setting sample ploidy to 1, and GenotypeGVCFs,
151 including non-variant sites in output VCF files. We included variant sites with a minimum depth of 5X
152 and a minimum variant quality score 20 and constructed consensus sequences with bcftools consensus⁵²,
153 excluding indels. We flagged SNPs in previously defined repetitive regions (PPE and PE-PGRS genes,
154 phages, insertion sequences and repeats longer than 50 bp)⁵³ and excluded these variants in figures and
155 statistics except when otherwise noted. We identified sub-lineage and evidence of mixed infection with
156 TBProfiler v.4.2.0^{54,55}.

157 We constructed full-length consensus FASTA sequences from VCF files, setting missing
158 genotypes to missing, and used SNP-sites to extract a multiple alignment of internal variant sites only⁵⁶.
159 We used the R package *ape* to measure pairwise differences between samples (*dist.dna*,
160 pairwise.deletion=TRUE)⁵⁷. We selected a best fit substitution model with ModelFinder⁵⁸, implemented

161 in IQ-TREE multicore version 2.2.0⁵⁹, evaluating all models that included an ascertainment bias
162 correction for the use of an alignment of SNPs only. We then fit a maximum likelihood tree with IQ-
163 TREE, with 1000 ultrafast bootstrap replicates^{59,60} to visualize the location of household pairs in the
164 context of study-wide variation.

165 We filtered variants that had coverage higher or lower than two standard deviations from the
166 sample mean depth, reasoning that the extreme coverage was a result of incorrect mapping. We
167 considered minority variants as positions with two or more alleles each supported by at least 5X coverage
168 at the same position, at first, without filtering by minor allele frequency threshold. To examine the impact
169 of filtering approach on the informativeness of minority variation, we applied increasingly conservative
170 minor allele thresholds, from 0.05% to 50%. We quantified the number of per-sample minority variants;
171 the number of shared minority variants between household members, defined as sharing the same minor
172 allele call at the same position; and the number of shared minority variants between epidemiologically
173 unrelated pairs.

174 Following variant identification, all analyses were conducted in R version 4.2.2. All analysis
175 scripts are available on GitHub (<https://github.com/ksw9/mtb-within-host>).

176

177 **Role of the funding source**

178 The study sponsor played no role in study design; in the collection, analysis, and interpretation of data; in
179 the writing of the report; and in the decision to submit the paper for publication.

180

181 **Results**

182

183 *M. tuberculosis* variation observed in household transmission studies.

184 To characterize the epidemiological information held in within-host *M. tuberculosis* variation
185 present in routinely generated Illumina sequence data from cultured isolates, we reanalyzed sequence data
186 from three previously published *M. tuberculosis* transmission studies for which whole genome

187 sequencing data and epidemiological linkages were publicly available. Studies were from different
188 epidemiological settings and included (a) a household transmission study in Vitória, Brazil⁴⁰ (Colangeli et
189 al.), a retrospective population-based study of pediatric tuberculosis in British Columbia, Canada⁴¹
190 (Guthrie et al.), and (c) a retrospective population-based study in Oxfordshire, England¹¹ (Walker et al.).
191 Study design, sampling design, and culture and sequencing methods differed across studies (Table 1).

192 As reported in the original studies, we observed limited fixed variation between *M. tuberculosis*
193 consensus sequences from isolates collected within the same household or among isolates from patients
194 with epidemiological linkages compared to randomly selected pairs of sequences from the same population
195 (Fig. 1a). Consensus *M. tuberculosis* sequences from epidemiologically linked individuals were
196 phylogenetic nearest neighbors for each study (Fig. 1b). However, genetic distances between consensus
197 sequences often exceeded commonly used 5 and 12 SNP thresholds^{10,11} for classifying isolates as
198 potentially linked in transmission, with 44.4% (20/45) of household pairs not meeting a 5-SNP threshold
199 and 15.6% (7/45) of household pairs not meeting a 12-SNP threshold (Fig. 1a). Twenty-four percent
200 (11/45) of isolate pairs from epidemiologically linked individuals were within a genetic distance of 2
201 SNPs or less, underscoring that genomic distances alone may be limited in their resolution.

202

203 *Within-host variation observed in routine, culture-based M. tuberculosis sequencing data.*

204 We quantified minority variation within samples as the number of positions with a minor allele
205 above a frequency of a range of threshold values, as we were interested in tradeoffs between sensitivity
206 and specificity of variant detection. We detected limited, but measurable, minority variation above a 1%
207 minor allele frequency threshold, with a disproportionate number of minority variants occurring within
208 the PE/PPE genes (24.8%, 82.2%, and 80.1% of all minority variants, across the studies) (Fig. 2). We
209 found substantial differences in minority variation detected across studies with the Colangeli et al. study
210 (median: 160 minority variants, IQR:130-220) identifying a higher level of minority variation than both
211 the Guthrie et al. study (median: 3, IQR:1-8; Wilcoxon test, $p < 0.005$) and the Walker et al. study
212 (median: 2, IQR: 1-4, $p < 0.005$) (Table 3).

213 Most minority variants were in unique genomic locations and no minority variant was found in
214 more than 5 samples in a single study (Fig. S1). About half of minority variants were predicted to be
215 missense variants (50.0%; 964/1929) and only 1.3% (25/1929) minority variants were stop mutations,
216 which would generate a truncated protein. However, the 5 most common minority variants across all three
217 studies occurred in intergenic regions.

218 Median depth of coverage was significantly correlated with the total number of iSNVs detected
219 outside the PE/PPE genes for the Walker et al. study, though no association was identified in the
220 Colangeli et al. or Guthrie et al. studies (Fig. S2). Additionally, minor allele frequency was negatively
221 correlated with site depth of coverage in the Colangeli et al. and the Walker et al. studies, but not Guthrie
222 et al. (Fig. S2), potentially indicating that both culture method and sequencing depth were responsible for
223 the observed differences in recovered variation (Table 1).

224

225 *Signatures of transmission in within-host M. tuberculosis variation.*

226 To test whether within-host variation could be used to identify potential transmission linkages, we
227 quantified the number of shared minority variants passing quality, depth, and frequency thresholds
228 between each pair of samples in each study. Isolates from household pairs shared more minority variants
229 $\geq 1\%$ frequency and outside of PE/PPE genes than did randomly selected pairs of isolates in all three
230 studies (mean 98 shared minority variants vs. 10; 0.8 vs. 0.2; and 0.7 vs. 0.2, respectively; all $p < 0.001$,
231 Wilcoxon) (Table 2; Fig. 3). This effect rapidly declines as the definition of minority variant becomes
232 more stringent (Fig. S3). In each study, the distribution of shared minority variants differed significantly
233 between epidemiologically unlinked isolate pairs and epidemiologically linked pairs (Fig. 4a).

234 In a general linear model, shared within-host variation $\geq 1\%$ frequency and outside of PE/PPE
235 genes was significantly associated with household membership (OR: 1.51 [1.30,1.71] for one standard
236 deviation increase in shared minority variants. Genomic clustering, based on a standard 12-SNP
237 clustering distance thresholds, was also significantly associated with household membership (OR: 3,670
238 [1,160, 15,380]), with similar results when applying a 5-SNP clustering distance threshold. We measured

239 the performance of general linear models in classifying household pairs versus unlinked pairs with
240 receiver operator characteristic (ROC) curves. Including shared within-host variation improved the
241 accuracy of predictions in all three studies as compared to a model without within-host variation (AUC:
242 0.95 *versus* 0.92, 0.99 *versus* 0.95, and 0.93 *versus* 0.91) (Fig. 4b). A model including within-host
243 variation independently of consensus sequence-based clustering resulted in AUCs of 0.69, 0.64, and 0.64
244 for each study (Fig. 4b).

245 A major challenge in studies of pathogen variation and within-host variation, is distinguishing
246 true biological variation from errors introduced through sampling, sequencing, and bioinformatic
247 identification of variation in sequence data. To assess tradeoffs in sensitivity and specificity in minority
248 variant identification, we applied a series of increasingly conservative minor allele frequency thresholds,
249 filtering variants below a 0.05% to 50% frequency. Maximum AUC for predicting household membership
250 was 0.998 (minor allele frequency threshold: 2%) for the Colangeli et al. study, 0.996 (threshold: 5%) for
251 the Guthrie et al. study, and 0.94 (threshold: 5%) for the Walker et al. study (Fig. S4).

252 Among epidemiologically unlinked pairs, shared iSNVs declined significantly with increased
253 genetic distance between samples across all studies (Fig. S5). For household pairs, we did not find a
254 significant correlation between the genetic distance between isolate consensus sequences and number of
255 shared minority variants in the Colangeli et al. and Walker et al. studies (Fig. S5), suggesting that this
256 relationship may not be linear. While we did find a positive correlation between genetic distance and
257 shared iSNVs for the Guthrie et al. study, this was due to a single pair with a genetic distance of greater
258 than 20 SNPs.

259 Allele frequencies of shared minority variants $\geq 1\%$ frequency located outside of PE/PPE genes
260 were correlated between isolates from household pairs in Colangeli et al. (Pearson's $r=0.17$, $p<0.001$) and
261 Guthrie et al. ($r=0.94$, $p<0.001$), but not Walker et al. (Fig. S6). We predicted that sampling time might
262 impact recovery of shared minority alleles because of changes in allele frequency between the time of
263 sampling and time of transmission. Shared minority variation was negatively correlated with time
264 between collection of isolates from household index cases and household members, though the

265 association was not significant in the Colangeli et al. study, which reported time between sampling (Fig.
266 S7).

267

268 **Discussion**

269 To maximize the epidemiological information gleaned from the continuous evolution of *M.*
270 *tuberculosis*, approaches to leverage biological variation more fully are needed. Here, we found that (1)
271 within-host *M. tuberculosis* variation persists in sequence data from culture, (2) the magnitude of within-
272 host variation varies between and within studies and is impacted by methodological choices, and (3) *M.*
273 *tuberculosis* isolates from epidemiologically linked individuals share higher levels of variation than do
274 unlinked individuals and shared within-host variation improves predictions of epidemiological linkage.
275 Our results suggest that minority variation could contribute epidemiological information to transmission
276 inferences, improving inferences from consensus sequences, and that alternative approaches to culture-
277 based sequencing may further contribute to this observed epidemiological signal.

278 As sequencing has become more efficient and less expensive, pathogen genomic studies have
279 begun to describe previously uncharacterized levels of minority variation within individual hosts and
280 shared between transmission pairs. For example, *M. tuberculosis* within-host variation has been used to
281 reveal an undetected superspreader event^{25,26} in a single large outbreak in the Canadian Arctic. In another
282 study, Goig et al. observed minority variants that were shared between epidemiologically linked
283 individuals, and one example of isolates from a four-person transmission cluster that all shared a minority
284 variant at different allele frequencies³⁵. The existence of shared minority variants suggests that multiple
285 variants present in a donor's infection persist through transmission and are maintained within the recipient
286 through population changes and immune pressures. A similar observation has been made for other
287 pathogens—shared within-host diversity of SARS-CoV-2 has been used to improve phylogenetic and
288 transmission inferences in empirically collected and modeled sequence data^{61–63}. Recently developed
289 transmission inference approaches include pathogen within-host diversity to infer transmission events^{64–}
290 ⁶⁷, but have not yet been applied to *M. tuberculosis*, which is unique in its slow substitution rate and long

291 and variable periods of latent infection. Future work is needed to develop automated, user-friendly
292 pipelines for transmission and phylogenetic inference that include both fixed genomic differences and
293 within-host variation.

294 Our findings that within-host diversity persists through culture and is impacted by methodological
295 choices underscore the further work needed to optimize approaches for highly accurate identification of
296 within-host variation. Each step of generating sequence data, including clinical sampling, sample
297 preparation, sequencing, bioinformatic pipeline, may introduce a bottleneck and/or bias the variation
298 recovered. For example, our observation that minority variants are concentrated in PE/PPE genes,
299 highlights the need for testing whether long read sequencing or alternative mapping approaches can
300 improve the accuracy of variant identification in this region⁴⁶. Further, we found that increased
301 sequencing coverage and, potentially, culture approach, detect higher levels of within-host variation.

302 A major challenge in pathogen genomics, including studies of within-host pathogen variation, is
303 in distinguishing true biological variation from noise introduced by sequencing, bioinformatic, or other
304 errors. There are significant trade-offs between sensitivity and specificity in variant identification; often,
305 pathogen genomic approaches err on the side of specificity and impose conservative variant filters. Our
306 findings here and previously⁴⁶ suggest that for studying transmission linkages, including low frequency
307 minority variants may improve predictions of transmission linkage. However, it is likely that some of the
308 minority variants within individual samples and shared across samples are artefacts. For example, we
309 found that some unlinked pairs of isolates share minority variants, potentially errors or true variants
310 occurring at highly mutable sites (Fig. 3).

311 There are several limitations to our study. First, we conducted a re-analysis of previously
312 published sequence data from clinical *M. tuberculosis* samples. We therefore do not have information
313 about the true biological variation present within samples and cannot assess sensitivity and specificity of
314 variants identified using alternative approaches. To measure performance of hybrid capture and other
315 methods in recovering true within-host variation and the limit of detection of within-host variation,
316 experiments that directly compare recovery of minority variants in known strain mixtures are required.

317 Second, we found that one study found substantially higher within-host variation than the others, likely
318 reflecting large differences in study design and sample preparation (Table 1). The Colangeli et al. was a
319 prospective study, and included three loops of culture for DNA extractions, while the Guthrie et al. and
320 Walker et al. studies were retrospective and re-cultured isolates after frozen storage. This difference could
321 also reflect higher population-wide *M. tuberculosis* diversity circulating in a higher-incidence setting. It is
322 possible that other steps in *M. tuberculosis* sampling, sampling time (i.e. Fig. S5), culture, laboratory
323 preparation, or sequencing influenced recovered within-host variation; if these steps were not reported,
324 we were not able to include them in our models of within-host variation. For example, data on sequencing
325 run, a potential source of false shared variation, was not available. Third, we considered household
326 transmission pairs as our gold standard for transmission linkages. While the studies we included
327 employed additional filters to exclude household pairs unlikely to be epidemiologically linked, it is
328 possible that these pairs are misclassified. However, the impact of such misclassification would be to bias
329 our results towards the null finding that shared minority variants are not more likely to be found in
330 transmission pairs than unlinked pairs. Finally, we do not have access to sequencing replicates of the
331 same sputum culture or biological replicates of the same sputum to quantify the concordance of minority
332 variants across sequencing or biological replicates.

333 Our findings of within-host variation present in cultured *M. tuberculosis* samples suggests that
334 within-host *M. tuberculosis* variation may be able to augment routine transmission inferences. More
335 broadly, these finding suggests that assessing *M. tuberculosis* variation more broadly, including not only
336 within-host variants, but also genome-wide variants and indels may yield more information and improve
337 both transmission and phylogenetic inferences.

338

339 **Declaration of interest**

340 The authors report no conflict of interest.

341

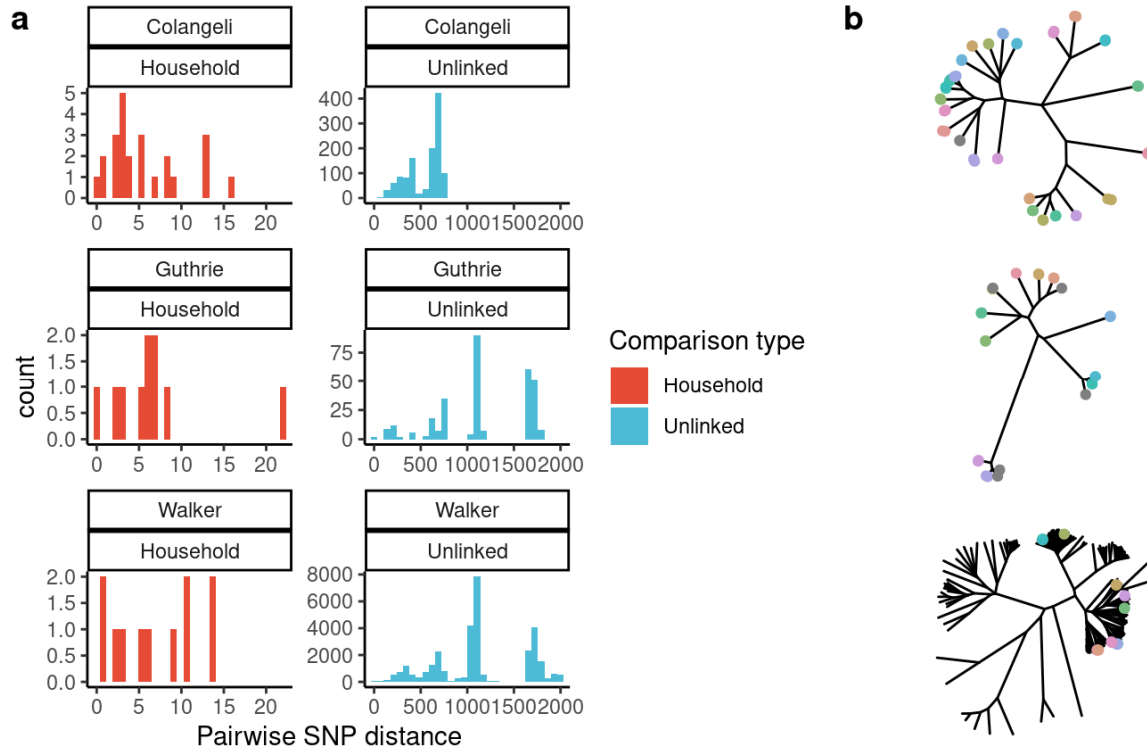
342

348 **Table 2. Measured within-host *M. tuberculosis* variation.** Per-sample and shared minority variants
349 across pairwise comparisons with different epidemiological linkages, including minority variants $\geq 1\%$
350 allele frequency, outside of the PE/PPE genes, and within an expected depth (defined in Methods).
351

Comparison type	mean	median	lower	upper
Colangeli				
Sample	170.00	160	130	220.0
Household	98.00	95	64	130.0
Unlinked	9.80	1	0	7.0
Guthrie				
Sample	5.80	3	1	8.0
Household	0.80	1	0	1.0
Unlinked	0.15	0	0	0.0
Walker				
Sample	7.10	2	1	4.0
Household	0.73	0	0	1.5
Unlinked	0.17	0	0	0.0

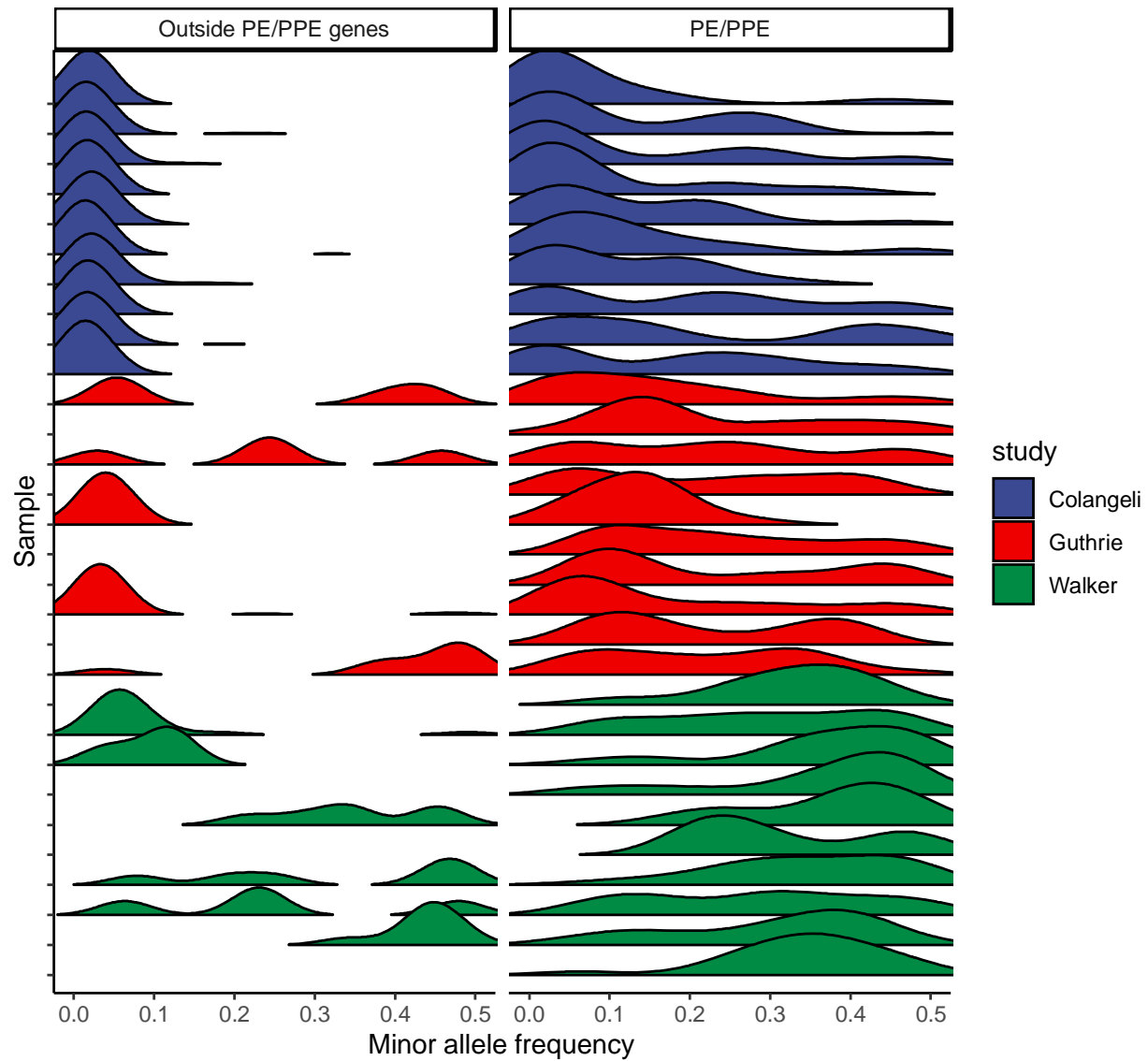
352
353

354 **Figure 1. *M. tuberculosis* consensus genomes are closely related but are not always predictive of**
355 **epidemiological linkage.** (a) Histograms indicate pairwise genetic distances between *M. tuberculosis*
356 consensus genomes, with facets indicating study and pairwise comparison type. (b) Phylogeny of
357 consensus sequences for each study, with branch tips colored to indicate samples from a single household
358 or with known epidemiologic links.
359



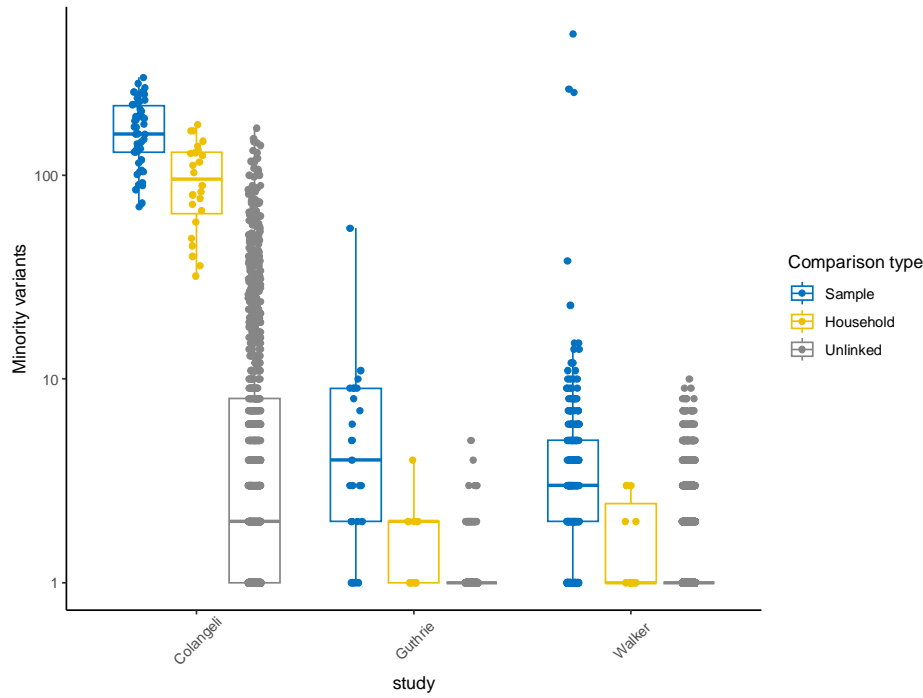
360
361

362 **Figure 2. Limited *M. tuberculosis* within-host diversity is recovered with culture-based Illumina**
363 **sequencing.** Ridgeline plot of the minor allele frequency distribution for five randomly selected samples
364 from each study, indicated by ridge color. Panels indicate genomic region: outside PE/PPE genes and
365 within PE/PPE genes.



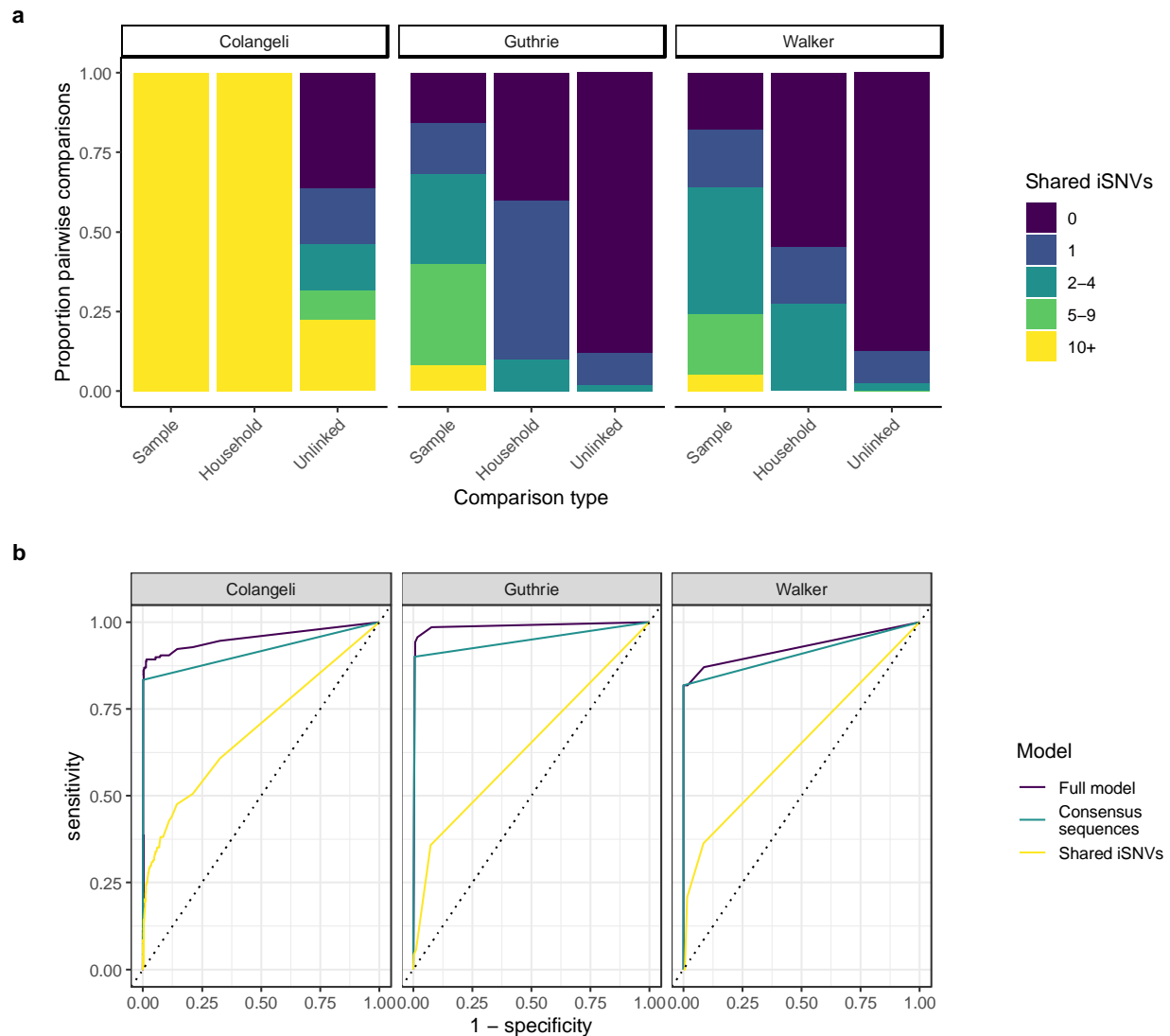
366
367

368 **Figure 3. Pairwise shared variants above a 1% minor allele frequency.** Boxplots of the number of
369 high-quality shared minority variants between sample pairs in three previously published *M. tuberculosis*
370 transmission studies (columns) with jittered points indicating pairwise observations. Colors indicate
371 comparison type: sample, within-host minority variants; household, minority variants shared between
372 household pairs; unlinked, minority variants shared between individuals in different households. Boxes
373 indicate group interquartile ranges and center lines indicate group medians.
374



375

376 **Figure 4. Shared minority variants contain information about household membership.** (a) Stacked
377 barplot showing the proportion of sample pairs across different levels of shared minority variants $\geq 1\%$
378 minor allele frequency threshold. Panels indicate study. (b) ROC curves showing sensitivity (true positive
379 rate) as a function $1 - \text{specificity}$ (true negative rate) for predicting household membership in general
380 linear models that include both shared iSNVs and consensus sequence-based clusters (Full model),
381 consensus sequence-based cluster only (Consensus sequences), and Shared iSNVs only (Shared iSNVs).
382 All models include study as a predictor.
383
384
385
386
387



388

389 **References**

- 390
- 391 1. Churchyard, G. *et al.* What We Know about Tuberculosis Transmission: An Overview. *Journal of*
392 *Infectious Diseases* **216**, S629–S635 (2017).
- 393 2. Auld, S. C. *et al.* Extensively drug-resistant tuberculosis in South Africa: Genomic evidence
394 supporting transmission in communities. *European Respiratory Journal* **52**, (2018).
- 395 3. Middelkoop, K. *et al.* Transmission of tuberculosis in a south African community with a high
396 prevalence of HIV infection. *Journal of Infectious Diseases* **211**, 53–61 (2015).
- 397 4. Andrews, J. R., Morrow, C., Walensky, R. P. & Wood, R. Integrating social contact and
398 environmental data in evaluating tuberculosis transmission in a South African township. *J Infect*
399 *Dis* **210**, 597–603 (2014).
- 400 5. Yates, T. A. *et al.* The transmission of Mycobacterium tuberculosis in high burden settings. *Lancet*
401 *Infect Dis* **16**, 227–238 (2016).
- 402 6. Andrews, J. R., Morrow, C. & Wood, R. Modeling the role of public transportation in sustaining
403 tuberculosis transmission in South Africa. *Am J Epidemiol* **177**, 556–561 (2013).
- 404 7. Ypma, R. J. F., Altes, H. K., Van Soolingen, D., Wallinga, J. & Marijn Van Ballegooijen, W. A
405 Sign of Superspreading in Tuberculosis: Highly Skewed Distribution of Genotypic Cluster Sizes.
406 *Source: Epidemiology* **24**, 395–400 (2013).
- 407 8. Gygli, S. M. *et al.* Prisons as ecological drivers of fitness-compensated multidrug-resistant
408 Mycobacterium tuberculosis. *Nat Med* **27**, 1171–1177 (2021).
- 409 9. Xu, Y. *et al.* High-resolution mapping of tuberculosis transmission: Whole genome sequencing
410 and phylogenetic modelling of a cohort from Valencia Region, Spain. *PLoS Med* **16**, (2019).
- 411 10. PHE. Tuberculosis in England: 2018 Presenting data to end of 2017. *Public Health England*
412 **Version 1.**, 173 (2018).
- 413 11. Walker, T. M. *et al.* Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK,
414 2007–12, with whole pathogen genome sequences: An observational study. *Lancet Respir Med* **2**,
415 285–292 (2014).
- 416 12. Bryant, J. M. *et al.* Inferring patient to patient transmission of Mycobacterium tuberculosis from
417 whole genome sequencing data. *BMC Infect Dis* **13**, 1–12 (2013).
- 418 13. Guerra-Assunção, J. *et al.* Large-scale whole genome sequencing of M. tuberculosis provides
419 insights into transmission in a high prevalence area. *Elife* **4**, 1–17 (2015).
- 420 14. Menardo, F., Duchêne, S., Brites, D. & Gagneux, S. The molecular clock of mycobacterium
421 tuberculosis. *PLoS Pathog* **15**, (2019).
- 422 15. Casali, N. *et al.* Evolution and transmission of drug-resistant tuberculosis in a Russian population.
423 *Nat Genet* **46**, 279–286 (2014).
- 424 16. Roetzer, A. *et al.* Whole Genome Sequencing versus Traditional Genotyping for Investigation of a
425 Mycobacterium tuberculosis Outbreak: A Longitudinal Molecular Epidemiological Study. *PLoS*
426 *Med* **10**, (2013).
- 427 17. Yang, C. *et al.* Internal migration and transmission dynamics of tuberculosis in Shanghai, China:
428 an epidemiological, spatial, genomic analysis. *Lancet Infect Dis* **18**, 788–795 (2018).
- 429 18. Borges, V. *et al.* Nosocomial Outbreak of SARS-CoV-2 in a “Non-COVID-19” Hospital Ward:
430 Virus Genome Sequencing as a Key Tool to Understand Cryptic Transmission. *Viruses* **13**, (2021).
- 431 19. Cohen, T. *et al.* Mixed-Strain Mycobacterium tuberculosis Infections and the Implications for
432 Tuberculosis Treatment and Control. *Clin Microbiol Rev* **25**, 708–719 (2012).
- 433 20. Cohen, T. *et al.* Within-host heterogeneity of mycobacterium tuberculosis infection is associated
434 with poor early treatment response: A prospective cohort study. *Journal of Infectious Diseases*
435 **213**, 1796–1799 (2016).
- 436 21. Pérez-Lago, L. *et al.* Revealing hidden clonal complexity in Mycobacterium tuberculosis infection
437 by qualitative and quantitative improvement of sampling. *Clinical Microbiology and Infection* **21**,
438 147.e1-147.e7 (2015).

- 439 22. Lieberman, T. D. *et al.* Genomic diversity in autopsy samples reveals within-host dissemination of
440 HIV-associated Mycobacterium tuberculosis. *Nat Med* **22**, 1470–1474 (2016).
- 441 23. Mann, B. C., Jacobson, K. R., Ghebrekristos, Y., Warren, R. M. & Farhat, M. R. Assessment and
442 validation of enrichment and target capture approaches to improve Mycobacterium tuberculosis
443 WGS from direct patient samples. *J Clin Microbiol* **61**, (2023).
- 444 24. Séraphin, M. N. *et al.* Direct transmission of within-host Mycobacterium tuberculosis diversity to
445 secondary cases can lead to variable between-host heterogeneity without de novo mutation: A
446 genomic investigation. *EBioMedicine* **47**, 293–300 (2019).
- 447 25. Lee, R. S., Proulx, J.-F. F., McIntosh, F., Behr, M. A. & Hanage, W. P. Previously undetected
448 super-spreading of mycobacterium tuberculosis revealed by deep sequencing. *Elife* **9**, (2020).
- 449 26. Martin, M. A., Lee, R. S., Cowley, L. A., Gardy, J. L. & Hanage, W. P. Within-host
450 mycobacterium tuberculosis diversity and its utility for inferences of transmission. *Microb Genom*
451 **4**, (2018).
- 452 27. Zetola, N. M. *et al.* Mixed Mycobacterium tuberculosis complex infections and false-negative
453 results for rifampin resistance by genexpert MTB/RIF are associated with poor clinical outcomes.
454 *J Clin Microbiol* **52**, 2422–2429 (2014).
- 455 28. Meehan, C. J. *et al.* Whole genome sequencing of Mycobacterium tuberculosis: current standards
456 and open issues. *Nat Rev Microbiol* **17**, 533–545 (2019).
- 457 29. McNerney, R. *et al.* Removing the bottleneck in whole genome sequencing of Mycobacterium
458 tuberculosis for rapid drug resistance analysis: a call to action. *International Journal of Infectious*
459 *Diseases* vol. 56 130–135 Preprint at <https://doi.org/10.1016/j.ijid.2016.11.422> (2017).
- 460 30. Martín, A., Herranz, M., Ruiz Serrano, M. J., Bouza, E. & García de Viedma, D. The clonal
461 composition of Mycobacterium tuberculosis in clinical specimens could be modified by culture.
462 *Tuberculosis* **90**, 201–207 (2010).
- 463 31. Plazzotta, G., Cohen, T. & Colijn, C. Magnitude and sources of bias in the detection of mixed
464 strain M. tuberculosis infection. *J Theor Biol* **368**, 67–73 (2015).
- 465 32. Walter, K. S. *et al.* Genomic variant-identification methods may alter mycobacterium tuberculosis
466 transmission inferences. *Microb Genom* **6**, 1–16 (2020).
- 467 33. Ates, L. S. New insights into the mycobacterial PE and PPE proteins provide a framework for
468 future research. *Mol Microbiol* 0–2 (2019) doi:10.1111/mmi.14409.
- 469 34. Phelan, J. E. *et al.* Recombination in pe/ppe genes contributes to genetic variation in
470 Mycobacterium tuberculosis lineages. *BMC Genomics* **17**, 1–12 (2016).
- 471 35. Goig, G. A. *et al.* Whole-genome sequencing of Mycobacterium tuberculosis directly from clinical
472 samples for high-resolution genomic epidemiology and drug resistance surveillance: an
473 observational study. *Lancet Microbe* **1**, e175–e183 (2020).
- 474 36. Brown, A. C. *et al.* Rapid whole-genome sequencing of mycobacterium tuberculosis isolates
475 directly from clinical samples. *J Clin Microbiol* **53**, 2230–2237 (2015).
- 476 37. Votintseva, A. A. *et al.* Same-day diagnostic and surveillance data for tuberculosis via whole-
477 genome sequencing of direct respiratory samples. *J Clin Microbiol* **55**, 1285–1298 (2017).
- 478 38. Nimmo, C. *et al.* Whole genome sequencing Mycobacterium tuberculosis directly from sputum
479 identifies more genetic diversity than sequencing from culture. *BMC Genomics* **20**, 1–9 (2019).
- 480 39. Doyle, R. M. *et al.* Direct Whole-Genome Sequencing of Sputum Accurately Identifies Drug-
481 Resistant Mycobacterium tuberculosis Faster than MGIT Culture Sequencing. *J Clin Microbiol* **56**,
482 666–684 (2018).
- 483 40. Colangeli, R. *et al.* Mycobacterium tuberculosis progresses through two phases of latent infection
484 in humans. *Nat Commun* **11**, (2020).
- 485 41. Guthrie, J. L. *et al.* Genotyping and Whole-Genome Sequencing to Identify Tuberculosis
486 Transmission to Pediatric Patients in British Columbia, Canada, 2005-2014. *J Infect Dis* **218**,
487 1155–1163 (2018).
- 488 42. Walter, K. S. *et al.* The role of prisons in disseminating tuberculosis in Brazil: A genomic
489 epidemiology study. *Lancet Regional Health - Americas* **9**, 100186 (2022).

- 490 43. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform.
491 *Bioinformatics* **25**, 1754–1760 (2009).
- 492 44. Van der Auwera, G. A. & O’Connor, B. *Genomics in the cloud* □: using Docker, GATK, and WDL
493 in Terra. *Genomics in the cloud* □: using Docker, GATK, and WDL in Terra (O’Reilly Media,
494 2020).
- 495 45. Van der Auwera, G. A. *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome
496 Analysis Toolkit Best Practices Pipeline. in *Current Protocols in Bioinformatics* vol. 43 11.10.1-
497 11.10.33 (John Wiley & Sons, Inc., 2013).
- 498 46. Walter, K. S. *et al.* Genomic variant-identification methods may alter Mycobacterium tuberculosis
499 transmission inferences. *Microb Genom* **6**, (2020).
- 500 47. Krueger, F. Trim Galore. Preprint at (2019).
- 501 48. Martin, M. Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads.
502 *EMBnet J* **17**, (2011).
- 503 49. Goig, G. A., Blanco, S., Garcia-Basteiro, A. L. & Comas, I. Contaminant DNA in bacterial
504 sequencing experiments is a major source of false genetic variability. *BMC Biol* **18**, (2020).
- 505 50. Wood, D. E. & Salzberg, S. L. Kraken: Ultrafast metagenomic sequence classification using exact
506 alignments. *Genome Biol* **15**, R46 (2014).
- 507 51. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of
508 NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
- 509 52. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, 1–4 (2021).
- 510 53. Brites, D. *et al.* A new phylogenetic framework for the animal-adapted mycobacterium
511 tuberculosis complex. *Front Microbiol* **9**, 2820 (2018).
- 512 54. Phelan, J. E. *et al.* Integrating informatics tools and portable sequencing technology for rapid
513 detection of resistance to anti-tuberculous drugs. *Genome Med* **11**, 41 (2019).
- 514 55. Coll, F. *et al.* Rapid determination of anti-tuberculosis drug resistance from whole-genome
515 sequences. *Genome Med* **7**, (2015).
- 516 56. Page, A. J. *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments.
517 *Microb Genom* **2**, 1–5 (2016).
- 518 57. Paradis, E. & Schliep, K. Ape 5.0: An environment for modern phylogenetics and evolutionary
519 analyses in R. *Bioinformatics* **35**, 526–528 (2019).
- 520 58. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermin, L. S.
521 modelfinder: fast model selection for accurate phylogenetic estimates. **14**, (2017).
- 522 59. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in
523 the Genomic Era. *Mol Biol Evol* **37**, 1530–1534 (2020).
- 524 60. Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving
525 the Ultrafast Bootstrap Approximation. *Mol Biol Evol* **35**, 518–522 (2018).
- 526 61. Torres Ortiz, A. *et al.* Within-host diversity improves phylogenetic and transmission
527 reconstruction of SARS-CoV-2 outbreaks. doi:10.7554/eLife.
- 528 62. Walter, K. S. *et al.* Challenges in Harnessing Shared Within-Host Severe Acute Respiratory
529 Syndrome Coronavirus 2 Variation for Transmission Inference. *Open Forum Infect Dis* **10**, (2023).
- 530 63. Siddle, K. J. *et al.* Transmission from vaccinated individuals in a large SARS-CoV-2 Delta variant
531 outbreak. *Cell* **185**, 485-492.e10 (2022).
- 532 64. De Maio, N., Wu, C. H. & Wilson, D. J. SCOTTI: Efficient Reconstruction of Transmission
533 within Outbreaks with the Structured Coalescent. *PLoS Comput Biol* **12**, e1005130 (2016).
- 534 65. De Maio, N., Worby, C. J., Wilson, D. J. & Stoesser, N. Bayesian reconstruction of transmission
535 within outbreaks using genomic variants. *PLoS Comput Biol* **14**, e1006117 (2018).
- 536 66. Wymant, C. *et al.* PHYLOSCANNER: Inferring transmission from within- and between-host
537 pathogen genetic diversity. *Mol Biol Evol* **35**, 719–733 (2018).
- 538 67. Worby, C. J., Lipsitch, M. & Hanage, W. P. Shared Genomic Variants: Identification of
539 Transmission Routes Using Pathogen Deep-Sequence Data. *Am J Epidemiol* **186**, 1209–1216
540 (2017).

