Mycobacterium tuberculosis complex lineage 5 exhibits high levels of within-lineage genomic diversity

C. N'Dira Sanoussi[1,2,3], Mireia Coscolla[4], Boatema Ofori-Anyinam[5], Isaac Darko Otchere[6], Martin Antonio[7], Stefan Niemann[8,9], Julian Parkhill[10,11], Simon Harris[10], Dorothy Yeboah-Manu[6], Sebastien Gagneux[12,13], Leen Rigouts[2,3], Dissou Affolabi[1], Bouke C. de Jong[2], Conor J. Meehan[2,14]

[1] Laboratoire de Référence des Mycobactéries, Cotonou, Benin

[2] Mycobacteriology Unit, Institute of Tropical Medicine, Antwerp, Belgium

[3] Department of Biomedical Sciences, University of Antwerp, Antwerp, Belgium

[4] I2SysBio, University of Valencia-FISABIO joint Unit, Valencia, Spain

[5] Center for Global Health Security and Diplomacy, Ottawa, Canada

[6] Noguchi Memorial Institute for Medical Research, University of Ghana, Legon, Accra, Ghana

[7] Medical Research Council Unit in The Gambia at the London School of Hygiene and Tropical Medicine, Fajara, The Gambia

[8] Research Center Borstel, Molecular and Experimental Mycobacteriology, Borstel, Germany

[9] German Center for Infection Research, partner site Borstel-Hamburg-Lübeck-Riems, Borstel Germany

[10] Wellcome Sanger Institute, United Kingdom

[11] Department of Veterinary Medicine, University of Cambridge, Cambridge, UK

[12] Swiss Tropical and Public Health Institute, Basel, Switzerland

[13] University of Basel, Basel, Switzerland

[14] School of Chemistry and Biosciences, University of Bradford, Bradford, UK

Abstract

Pathogens of the *Mycobacterium tuberculosis* complex (MTBC) are considered monomorphic, with little gene content variation between strains when compared with other bacteria. Nevertheless, several genotypic and phenotypic factors separate the different MTBC lineages (L), especially L5 and L6 (traditionally termed *Mycobacterium africanum*), from each other. However, genome variability and gene content especially of L5 and L6 strains have not been fully explored, but may be potentially important for pathobiology and current approaches for next generation sequencing (NGS) analysis of MBTC genomes.

Through genomic comparison of 208 L5 clinical isolates genomes (including 3 completed genomes and 205 Illumina NGS datasets) and H37Rv, we identified multiple genes differentially present or absent between H37Rv and L5 strains. Additionally, considerable gene content variability was found between L5 strains. Several of the unique L5 genes contain significant diversity in pairwise L5 strains comparison, thus, providing additional discriminatory power e.g. for genome-based transmission analysis that would be missed in the current H37Rv-centric mapping approach.

In conclusion, our data show that using H37Rv as reference genome results in missing SNPs in genes unique for L5 strains. This potentially leads to an underestimation of the diversity present in the genome of L5 strains. As such, a full capture of the gene diversity e.g. for high resolution outbreak analysis requires a variation of the single reference genome mapping

approach currently used in most NGS data analysis pipelines. Moreover, the high within-lineage gene content variability suggests that the pan-genome of *M. tuberculosis* may be larger than previously thought, implying a concatenated or reference-free genome assembly approach may be needed for particular questions.

Keywords: gene content, *M. africanum*, Lineage 5, H37Rv, reference genome, genomic diversity, within-lineage variability

Data summary

All supporting data and protocols have been provided within the article or through supplementary data files.

Introduction

Tuberculosis (TB) is caused by pathogenic bacteria of the *Mycobacterium tuberculosis* complex (MTBC) that consists of 8 human-adapted lineages and several animal-adapted lineages (Mireia Coscolla and Gagneux 2014; Ngabonziza et al. 2020). This group is highly clonal with no horizontal gene transfer (Boritsch et al. 2016). Strains of particular lineages primarily are defined by large sequence polymorphisms (the presence or deletion of genomic regions) such as the TbD1 region (MTBC specific deletion 1) and other regions of difference (RD) (Brosch et al. 2002; Gordon et al. 1999; Behr et al. 1999), and signature single nucleotide polymorphisms (SNPs)(Coll et al. 2014). Broadly, the human-adapted lineages (L1-8) are subdivided into ancestral lineages (L1, L5, L6 and L8), an intermediate lineage (L7) and modern lineages (L2, L3, L4)(Coscolla and Gagneux 2014; Ngabonziza et al. 2020) based on the presence/absence of specific RDs, especially TbD1 (Firdessa et al. 2013; Brosch et al. 2002;

67 Nebenzahl-Guimaraes et al. 2016). L5 and L6 are also called *M. tuberculosis* var. africanum (or
68 historically *M. africanum* West-african 1 and 2 respectively (Riojas et al. 2018; Niemann et al.
69 2004)) and are restricted to West Africa where they cause up to 40% of human TB (de Jong,
70 Antonio, and Gagneux 2010). Reasons for the geographical restriction of L5 and L6 remain
71 unclear, however, adaptation to particular human subpopulation has been
72 suggested(Intemann et al. 2009; Thye et al. 2011; Comas et al. 2013; Gagneux et al. 2006).

73 Several phenotypic and genotypic features separate L5 and L6 from the other human-adapted
74 lineages. Some TB diagnostics have a lower performance for L5 and L6 strains, compared to
75 other MTBC lineages (Sanoussi et al. 2018; Ofori-Anyinam et al. 2016) and these lineages are
76 less likely to grow in culture (Sanoussi et al. 2017), with dysgonic appearance on solid medium
77 (Leao et al. 2004; Pattyn et al. 1970; Sanoussi et al. 2018; Niemann et al. 2004). Mutations in
78 candidate genes essential for growth in culture were identified for L6 strains (Gehre et al.
79 2013), yet for L5 strains the reasons for the difficulty in growth remain unclear but may be
80 related to the presence of the TbD1 region and the absence of RD9 (Brosch et al. 2002). L6
81 strains and those of the closely related animal strains/lineages such as *M. bovis* are reported
82 to be less virulent than those from other human-adapted MTBC lineages in population-based
83 studies (deJong et al. 2008; Magnus 1966; Reiling et al. 2013) and in genome studies (Gonzalo-
84 Asensio et al. 2014; Ofori-Anyinam et al. 2020). Infection with an L6 strain progresses slowly
85 to TB disease, and is associated with impaired immunity in some settings but not all (e.g. HIV
86 infection (de Jong et al. 2008)). In contrast, little is known about the genomics, virulence and
87 disease progression of L5 strains.

88 Whole genome sequencing (WGS) based on next generation sequencing (NGS) of MTBC
89 strains often involves data analysis by mapping of short sequence reads to a reference

90 genome, usually H37Rv (Cole et al. 1998; Camus et al. 2019; Meehan et al. 2019) or a
91 reconstructed ancestor with the same gene content as H37Rv (Comas et al. 2013; Goig et al.
92 2019, 2020). However, since H37Rv is a L4 strain, it might not be representative of other MTBC
93 lineages, including L5 strains. Thus, if H37Rv is used as basis for genome analysis, several
94 genes or larger genomic regions may be missed resulting in an underestimation of genome
95 diversity among strains of other lineages that may, however, provide additional information
96 e.g. for transmission analysis.

97 The members of the MTBC evolved from an environmental organism to an obligate pathogen
98 through genome reduction and acquisition of new genes (Gagneux 2018) and it is known that
99 some differences in gene content exist between lineages (Meehan et al. 2019; Bifani et al.
100 2000; Periwal et al. 2015; Kato-Maeda et al. 2001; O'Toole and Gautam 2017). Furthermore,
101 it was reported that genes had a higher genetic diversity among L6 strains compared to L5
102 ones (Darko Otchere et al. 2018), but the difference in gene content between these two
103 lineages was not investigated. Also, little is known about the gene differences between L4 and
104 L5 and the potential limitations this may impose for in-depth analysis of L5 genome studies
105 (e.g. sub-lineage detection and transmission tracking). Similarly, little is known about within-
106 lineage diversity in terms of gene content.

107 To address these questions, we assessed the gene content diversity of L5 by analyzing in the
108 context of whole genome sequencing and reference selection. To this end, we analyzed 208
109 genomes of L5 strains including three completed genomes and compared them to H37Rv and
110 closely related lineages (L6, *M. tuberculosis* var. bovis (hereafter called *M. bovis*)(Riojas et al.
111 2018)). Main focus was to determine particularities in the genomes of L5 strains and to define

112 the level of within-lineage gene content differences among strains of this lineage of the *M.*
113 *tuberculosis* complex.

114

115

116 ## Materials and Methods

117 ### Genomes

118 **Complete genomes (PacBio, completed long reads)**. Three completed L5 genomes from
119 clinical isolates, all sequenced with the PacBio SMRT technology, were analyzed. One genome
120 was from a Benin isolate (PcbL5Ben; sequenced in this study), one from an isolate from The
121 Gambia (PcbL5Gbia, WBB1453_11-00429-1) (Phelan et al. 2018), and one from Nigeria
122 (PcbL5Nig, WBB1454_IB091-1) (Phelan et al. 2018). These genomes are also from disparate
123 parts of the L5 diversity, representing a broad range of L5 sub-lineages, as were recently
124 defined (Coscolla et al. 2020). The previously published reference/complete genomes H37Rv
125 (L4) (NC_000962.3)(Cole et al. 1998; Camus et al. 2019; Lew et al. 2011a), L6 (GM041182,
126 Genbank accession n°: FR878060.1, GCF_0001593225.1_ASM159322v1) (Bentley et al. 2012)
127 and *M. bovis* (AF2122/97, accession: LT708304.1) (Malone et al. 2017) were also included.

128 **Whole genomic DNA extraction**. Genomic DNA extraction was performed on growth from
129 fresh Löwenstein-Jensen slants using the semi-automated Maxwell 16 Cell DNA purification
130 kit in the Maxwell 16 machine, or from the late exponential phase of growth in 7H9 medium,
131 using the CTAB method (Belisle and Sonnenberg 1998; Coscolla et al. 2020)

132 **Assembly and annotation of PcbL5Ben.** The Benin Pacbio genome was assembled using the
133 in-house Sanger pipeline and checked for sufficient quality and coverage. The completed

134 genome is available from xxx. The genome sequences of all three completed PacBio genomes
135 were annotated using Prokka-1.12 (Seemann 2014) based on the reference genome H37Rv
136 annotation.

137 **Gene presence-absence analysis**

138 Gene content differences were assessed with an all-vs-all BLASTn approach, using BLAST+
139 version+2.8.1 (Altschul et al. 1990; Camacho et al. 2009) (Fig.1). For a specific genome-
140 genome comparison, the following procedure was used: the gene sequences of genome 1 (ffn
141 file; coding regions only) were compared to those of genome 2 using the default cut-offs to
142 look for any homology for each gene. Those genes found in genome 1 and not in genome 2
143 were then compared to the completed PacBio genome (fna file; coding and non-coding
144 regions) of genome 2 to look for pseudogenes using BLASTn. These pseudogenes were then
145 confirmed using tBLASTn of the genome 1 protein sequences (faa file) compared to the
146 completed PacBio genome (fna file) of genome 2. This procedure was used to compare all
147 three L5 completed PacBio genomes to each other as well as each to H37Rv. Those genes
148 present or absent in H37Rv (or pseudogenes in either) were compared in a similar manner to
149 the L6 and *M. bovis* reference strains to determine if these genes/pseudogenes are L5-specific
150 (i.e. present in L5, but not in H37Rv, L6 and *M. bovis*).

151 The ffn files of H37Rv, the three L5 completed PacBio genomes and the L6 complete genomes,
152 were also aligned using progressive Mauve (ver 20150226)(Darling et al. 2004) to identify
153 rearrangements and examine synteny.

154 **L5 Illumina (short reads) genomes.** In total 205 L5 isolates genomes sequenced on the
155 Illumina platform from various countries were included in the study. These genomes were
156 derived from a larger study on genetic diversity of L5 and L6 (Coscolla et al. 2020) with

157 reduced redundancy isolates removed (i.e. one representative retained for those that were
158 extremely closely related). The isolates (genomes) originated from West, South, East and
159 Central Africa, but primarily (n=157) from two regions within West Africa: the Western part
160 of West Africa (including The Gambia, Senegal, Mauritania, Sierra Leone, Ivory Coast, Liberia,
161 Guinea and Mali) and the Eastern part of West Africa (including Ghana, Nigeria, Benin Niger
162 and Burkina Faso) (Coscolla et al. 2020). In general, L5 and L6 are geographically restricted to
163 West and Central-Africa, making this L5 genome selection representative of all the
164 geographical regions affected.

165 **Mapping of L5 Illumina reads to H37Rv and L5 completed PacBio genomes.**

166 Raw reads (fastq files) of the genomes were mapped respectively to H37Rv and each of the 3
167 L5 completed PacBio genomes, using the MTBseq pipeline (Kohl et al. 2018). The depth
168 mapping coverage of the samples in the clinical WGS data Illumina genomes against the
169 reference genomes (percent read mapping, unambiguous coverage mean) was compared
170 between the three L5 completed PacBio genomes and between each L5 completed PacBio
171 genome and H37Rv. Mapping statistics parameters such as percent unambiguous total base,
172 uncovered bases, SNPs, deletions, insertions, substitutions, percent genes mapped to
173 reference, were also compared using each L5 completed PacBio genome or H37Rv as a
174 reference.

175 **Checking unique versus missing genes in Illumina L5 genomes, and SNPs in those confirmed**
176 **as L5 specific.** The genes found present or absent in L5 based on genome comparisons of the
177 three L5 completed PacBio genomes with H37Rv were checked for their expected presence
178 or absence in the L5 Illumina genomes. Using the Position Tables produced by the MTBseq
179 pipeline, a gene was considered absent if 95% of its position in the genome had less than 8

180 reads covering them. From this data, a gene presence/absence matrix was generated for L5
181 Illumina genomes mapped to H37Rv and each of the three L5 completed PacBio genomes.
182 Genes found to be L5-specific (present in completed PacBio and Illumina genomes) were also
183 checked for SNPs in these genes using each of the L5 completed PacBio genomes as a
184 reference. The script used for undertaking this analysis can be found at
185 https://github.com/conmeehan/pathophy.

186 **Calculating the effect of reference genome selection on pairwise SNP distances**

187 Transmission analysis of MTBC isolates often involves clustering isolates together based on
188 specific SNP cut-offs (Meehan et al. 2019, 2018; Walker et al. 2013). In order to assess
189 whether the selection of reference genome would alter the SNP distances between samples,
190 and thus potentially change transmission cluster estimations, we calculated the SNP distance
191 difference between samples when using different reference genomes. For example, if Sample
192 A and Sample B are 10 SNPs apart using H37Rv as the reference and are 8SNPs apart using
193 PcbL5Ben as the reference, the SNP distance difference between these reference genomes
194 for these samples is 2 SNPs.

195 These SNP distance differences were calculated by comparing SNP alignments of the Illumina-
196 sequenced genomes resulting from mapping to each of the four reference genomes (H37Rv
197 and 3 L5-PacBio). The Amend function of MTBseq does this automatically for H37Rv, including
198 masking of repetitive regions, accounting for 10% of the genome (Cole et al. 1998), and
199 exclusion of columns with 95% ambiguous calls. To undertake this for the three L5 reference
200 genomes, repetitive regions were first determined from the annotations of the genes. All
201 genes whose description contained one of the following words was excluded: integrase, PE
202 family, PE-PGRS, phage, transposase. SNP alignments were then created using MTBseq as was

203 done for H37Rv. Pairwise distance matrices, one for each alignment based on each reference
204 genome, were created using a python script which can be found at
205 https://github.com/conmeehan/pathophy. SNP distance differences were calculated for each
206 sample pair between these matrices in R version 3.6.1 (R Core Team 2019) with the lower
207 triangle of the matrix extracted using gdata (Warnes et al. 2017) to ensure pairwise distances
208 were not counted twice. Histograms of pairwise SNP distance differences between reference
209 genomes were then plotted using ggplot2(Gómez-Rubio 2017).

210 **Determination of putative function of genes differentially present or absent in the L5**
211 **completed PacBio genomes**

212 To find the (putative) function of genes present or absent only in L5 strains, and genes present
213 in L5 strains but pseudogenes in H37Rv/L6/*M. bovis* or vice-versa, the fasta sequence of the
214 genes were searched against the NR database of NCBI using BLASTX-2.8.1+ (Altschul et al.
215 1997) as well as against the Tuberculist database (http://tuberculist.epfl.ch)(Lew et al.
216 2011b), Mycobrowser (https://mycobrowser.epfl.ch/) (Kapopoulou, Lew, and Cole 2011) and
217 literature. Furthermore, the gene function group/class was found using the COG database
218 (Tatusov et al. 2000) and Mycobrowser.

219

220 ## Results

221 **Comparative analysis of the completed L5 PacBio genomes and other lineages reveal**
222 **differences in gene content**

223 The number of genes, including paralogs, in the completed L5 PacBio genomes was: 4189
224 genes in the PcbL5Ben genome, 4162 in PcbL5Gbia, and 4134 in PcbL5Nig versus 4126 in

225 H37Rv, 4126 in the reference L6 genome and 4059 in the reference *M. bovis* genome. Hence,
226 the maximum gene count difference among the three L5 isolates from this study was 55
227 genes, among the human-adapted lineages (H37Rv, L5, L6) was 63 genes, and 130 among the
228 human- and animal adapted lineage (H37Rv, L5, L6 and *M. bovis*).

229 The visualization of the structure of the genomes showed that all three L5 genomes had a
230 region (we called region 1) that is absent in H37Rv (gap between 1683331-1716665 on Fig.2).
231 Furthermore, the Nigerian L5 completed PacBio genome (PcbL5Nig) was missing an additional
232 region (we called region 2, ~33334 bp, region between 1966666-2000000 in Fig.2) present in
233 the Benin and Gambian L5 genomes (Fig.2) and also present in H37Rv.

234 The Benin genome contained 5 genes (randomly spread in the genome) that were neither
235 present in the Nigerian or Gambian genomes, while the others contained respectively 1 and
236 2 unique genes. The Benin and Gambian strains each contained 33 genes that were not
237 present in the Nigerian strain (Fig.3, Table 1, Table S1). That's the region 2 mentioned above.
238 Interestingly, 32 of these genes were also present in H37Rv (Fig.4, Table 1, Table S1). Those
239 32 genes- absent in the Nigerian L5 strain only- were sequentially contiguous and formed
240 three blocks of 18, 12 and 2 genes (Table S1). The blocks of 18 and 12 genes were separated
241 by one gene and represented the region (mentioned in the section above, Fig 2) found missing
242 in the Nigerian L5 genome after the comparison of the structure of the 3 L5 completed PacBio
243 genomes. Ten genes (region 1 mentioned above) that were shared by the three L5 genomes
244 were absent in H37Rv, while 9 genes (spread across the genome) present in H37Rv were not
245 present in any of the three L5 genomes. Two genes (*Rv2073c*, *Rv2074*) of those 9 genes were
246 only present in H37Rv but completely absent in the L5, L6 and *M. bovis* complete genomes.

247 We qualified as 'suspected pseudogenes' any gene sequences that were not found in the ffn
248 file (sequences for coding regions) of a genome but found in its complete chromosome
249 sequence (fna) file (coding and non-coding regions of the genome sequence) that are not yet
250 confirmed as pseudogenes as described above in the methods. Six of the genes shared by the
251 three L5 completed PacBio genomes were confirmed pseudogenes in H37Rv (Table S2). Three
252 genes present in H37Rv were confirmed to be pseudogenes in the three L5 completed PacBio
253 genomes (Table S3).

254 **Gene presence/absence and related SNPs in lineage specific genes in a wider set of clinical**
255 **isolates**

256 The mapping estimates of the Illumina genomes are presented in Table 2. Mapping quality
257 and coverage against a completed L5 reference was superior to the H37Rv reference
258 approach (Table 2, Fig.5), as expected. Using the completed PacBio genomes from Benin and
259 The Gambia had similar mapping estimates that were better than those of the Nigerian one,
260 likely due to the large gene deletion in this genome. The genes specific to PcbL5Ben (5),
261 PcbL5Gbia (2), PcbL5Nig (1) were each found at various rates among the 205 L5 Illumina
262 genomes (77.6%-97.1%, 159-199/205, Table 3).

263 Interestingly, 2.9% of the Illumina L5 genomes (6/205) had similar patterns of large gene loss
264 as the Nigerian completed PacBio genome as they missed 30 of the 33 genes present in the
265 Benin and Gambian genomes and H37Rv. These six PcbL5Nig-like Illumina L5 genomes formed
266 a monophyletic group, within the L5 clade (Fig. 6), suggesting a single loss of these gene
267 clusters, although those six L5 genomes originated from several different countries including
268 Benin and Ghana. These two blocks contain genes whose annotations include *mutB*, *mazE4*,
269 *mazF4* and others with various putative functions (Table S4).

270 Four of the ten genes present in the three completed L5 PacBio genomes but absent in H37Rv
271 were found in all the 205 L5 Illumina genomes while the other were found in variable amounts
272 81-98% (165-201) (Table S5). These four genes include: *Mb2048c*, PE35, a hypothetical
273 protein- possibly an IS*256* transposase, and a hypothetical protein possibly related to the
274 CAAX conserved domain. Two of these four genes were present in L5 genomes only (PE35 and
275 hypothetical protein (possibly CAAX conserved domain)) while the remaining two (*Mb2048c*
276 and hypothetical protein (possibly IS256 transposase)) were found in L6 and *M. bovis* as well
277 (Table 4). Importantly for phylogenetic purposes, SNPs were detected in 1.5-3.4% of the 205
278 L5 Illumina genomes in all four genes (Table S5). Predicted functions for all L5-specific genes
279 are listed in table S5.

280 Five (*Rv1977*, *Rv1979c*, *Rv1993c*, *Rv1995*, *Rv2073c*) of the 9 genes present in H37Rv and
281 absent in the three completed L5 PacBio genomes were confirmed completely absent in all
282 the 205 L5 Illumina genomes. Four of these genes (*Rv1977*, *Rv1979c*, *Rv1993c*, *Rv1995*) were
283 absent in L5 strains only (i.e. present in L6 and *M. bovis* reference genomes) while the fifth
284 (*Rv2073c*) was absent in L6 and *M. bovis* as well (Table S6). The four other genes absent from
285 the completed PacBio genomes were present in a minority of the L5 Illumina genomes (Table
286 S6, Fig. 5). Predicted functions for all nine genes are listed in table S6.

287 **Effect of reference genome choice on between-isolate SNP distances**

288 We then assessed to what extent the difference in gene content between reference genomes
289 could affect genome analyses in transmission studies, when H37Rv is used as reference versus
290 a complete L5 genome as reference. Comparison of pairwise SNP distance matrices derived
291 from each of the four genomes showed that the difference in SNP distance between samples
292 varied widely depending on the reference genome used. When comparing H37Rv to each of

293 the PacBio L5 genomes, these differences ranged from 0 to 31 SNPs, with a mean of 10 SNPs
294 (Figure S1). This means that, on average, the calculated SNP distance between any two L5
295 samples differed by 10 SNPs as a consequence of changes in the selected reference genome.
296 When comparing the 3 PacBio L5 genomes to one another, the difference in SNP distance was
297 less affected, with differences ranging from 0 to 31 SNPs with a mean of 3 SNPs (Figure S2-
298 A,B,C). However, even such smaller differences in SNP distance between samples would still
299 potentially affect transmission studies. These results imply that transmission analyses cut-offs
300 could be greatly affected by the choice of reference genome, especially when using a non-L5
301 reference such as H37Rv.

302

303 **Discussion**

304 Comparison of completed PacBio and Illumina short-reads genomes from clinical isolates
305 (multi-country collection) revealed significant amounts of gene content diversity both
306 between L5 isolates and other lineages and within L5 itself.

307 We found that four genes, present in all the L5 isolates of our multi-country collection, were
308 absent in the H37Rv genome. Furthermore, those genes contain SNPs that can be used to
309 identify additional diversity between L5 strains. This has implications when using precise SNP
310 cut-offs to define transmission clusters (Walker et al. 2013; Meehan et al. 2018) as SNPs in
311 these genes would be missed with H37Rv mapping, resulting in underestimation of genome
312 diversity of L5 strains. Our findings suggest, that using an L5 genome as a reference would
313 likely increase the resolution of L5 genome analyses over the H37Rv approach and can
314 potentially increase strains discrimination for transmission analysis.

315 Our findings are in line with other reports indicating that some genes present in genomes of
316 clinical MTBC strains were absent in the H37Rv genome (Periwal et al. 2015; O'Toole and
317 Gautam 2017), and with other recommendations to use additional reference genomes
318 different to H37Rv (Norman et al. 2019; O'Toole and Gautam 2017; Lee et al. 2020). Although
319 Lee *et al*. (Lee and Behr 2016) - after comparing various lineages as reference genome for L4
320 genome analysis - concluded that there is no need to use a lineage-specific reference genome,
321 their observation was only based on the analysis of L4 clinical isolates and focused on SNPs
322 and short indels, not larger gene deletions or SNPs within these regions. In contrast, our
323 findings indicate that mapping of NGS data from L5 strains to a modern (L4) reference will
324 have an impact, in terms of both reference genome coverage and coverage of lineage-specific
325 genes.

326 Although the use of an L5 genome as reference would have many benefits over H37Rv for
327 particular study questions, several gene content differences were still observed within the L5
328 lineage. Each of the three L5 PacBio genomes had genes unshared with the two others that
329 were all found in L5 Illumina genomes (thus excluding exogenous contamination). The
330 absence of 30 genes in the Nigerian L5 completed PacBio genome (but shared by the Benin
331 and Gambian long read genomes), also found in six L5 genomes from different countries,
332 confirmed the large deletions in the PcbL5Nig genome and showed that it is representative
333 of some L5 strains in some countries. Deletions in the genomes of strains of MTBC lineages
334 were previously reported (Tsolaki et al. 2005; Gagneux et al. 2006; Kato-Maeda et al. 2001).
335 The fact that, on the phylogenetic tree the six isolates formed a monophyletic group of the
336 L5.3 sub-lineage (Coscolla et al. 2020) suggests that the loss of these 30 genes is a marker of
337 the L5.3 sub-lineage. Thus, we find that regions of difference (RDs) are present both between
338 lineages and between sub-lineages.

339 The classification of the L5-specific genes into functional groups based on Mycobrowser and
340 Tuberculist databases (Table 4) revealed a variety of functions including: a PE/PPE protein
341 family (possibly PE35) which may be linked to antigenic variation (NCBI annotation) and
342 virulence (STRING database annotation (Szklarczyk et al. 2017)); two ABC-transporters; an
343 IS*256*-like transposase, whose homology (IS256) is a marker of specific virulent lineages of
344 Staphylococci (Murugesan et al. 2018; Gu et al. 2005); and a gene with a CAAX conserved
345 domain, implicated in post-translation modification or signal transduction (Gao, Liao, and
346 Yang 2009a, 2009b). Only one (*Rv1979c*) of the four genes absent in only L5 strains had a
347 predicted function, labelled as "cell wall and cell processes" (Table S6). While *Rv1979c* is
348 associated with clofazimine and bedaquiline resistance (Zhang et al. 2015; Ghodousi et al.
349 2019), two of the drugs used for the treatment of RIF-resistant/MDR-TB (WHO 2016, 2019),
350 minimal inhibitory concentration testing of clofazimine in five L5 strains did not find the
351 deletion conferred resistance (Merker et al. 2020), further studies and protein function
352 discovery is needed to investigate the consequences of the absence of those genes in L5.

353 The genes *Rv1978* and *Rv1994c* present in H37Rv and the completed reference genomes of
354 L6 and *M. bovis*, were absent in the 3 completed Pacbio genomes of L5 strains, yet present in
355 few L5 isolates Illumina genomes (3/205, 1.5% for *Rv1978*; 1/205, 0.5% for *Rv1994c*). *Rv1978*
356 is required for bacterial survival in macrophages, and non-essential for *in vitro* growth of
357 H37Rv (Rengarajan, Bloom, and Rubin 2005). *Rv1994c* is involved in the regulation and
358 transport (efflux) of toxic metals, especially copper, which is toxic in excess (Samanovic et al.
359 2012) and may hamper *in vitro* growth, and survival during the chronic phase of infection,
360 similar to the effect of disruption of *csoR* (Marcus et al. 2016; Rowland and Niederweis 2012;
361 Ward, Hoye, and Talaat 2008; Cavet et al. 2003) (Table S6).

362 The functional annotation for the H37Rv-specific genes *Rv2073c* (short chain dehydrogenase)
363 and *Rv2074* (pyridoxamine-5-phosphate oxydase: vitamin B6 (pyridoxine)) absent in L5, L6
364 and *M. bovis* strains is "intermediary metabolism and respiration". Those two genes are part
365 of the RD9 region, which confirmed their absence in L5, L6 and *M. bovis* that have that region
366 deleted (Gordon et al. 1999). *Rv2073c* is an NAD(P)-dependent oxidoreductase (NCBI) that
367 catalyzes a wide range of reactions and is involved in redox sensor mechanisms (Kavanagh et
368 al. 2008; Sellés Vidal et al. 2018). *Rv2074* was previously thought to be a pyridoxine (vitamin
369 B6) oxidase, but is now known to be a F420 dependent biliverdin reductase, a cofactor of
370 vitamin B6 synthesis (UniProt) (Ahmed et al. 2016; Selengut and Haft 2010). Vitamin B6 is
371 essential for survival and virulence of *M. tuberculosis* (Dick et al. 2010) and its cofactor (F420
372 dependent biliverdin reductase) is implicated in immune-evasive mechanisms to allow
373 bacterial persistence (Ahmed et al. 2016; Selengut and Haft 2010)(Table S6). Recently, it has
374 been reported that the synthesis of F420 might be stimulated by phosphoenolpyruvate
375 (Grinter et al. 2020), which is the precursor to pyruvate, a supplement often added to culture
376 medium to improve the *in vitro* growth of L5, L6 and *M. bovis*. Furthermore, there is a
377 suggestion that F420 is needed for the activation of the antituberculosis drugs pretomanid
378 and delamanid (Grinter et al. 2020).

379 The absence of these genes in most L5 isolates suggests that L5 strains would be less likely to
380 survive in macrophages (*Rv1978*), have reduced growth *in vitro* (*Rv1994c* , as previously found
381 by Sanoussi et al (Sanoussi et al. 2017)), be less immune-evasive, less persistent and less
382 virulent (*Rv2074*) than L4 strains (at least H37Rv; as previously suggested for L5 and L6 strains
383 (Coscolla and Gagneux 2014; Reiling et al. 2013)). In addition, the absence of the gene *Rv2074*
384 in most L5 (and in the complete genomes of L6 and *M. bovis* as well) suggests that L6 and *M.*
385 *bovis* strains would also be less likely to be immune-evasive, to be persistent and to survive

386 than their L4 counterparts. Despite the presence of vitamin B6 in boiled egg (thus Lowenstein-
387 Jensen medium), its quantity in the L.J medium is probably insufficient to allow a high yield of
388 L5, L6 and *M. bovis* growth in culture (survival). Further studies, including vitamin B6
389 supplementation, could investigate the consequences of the absence of those genes in most
390 L5 strains (and L6 and *M. bovis* for *Rv2074*).

391 The within-lineage differences in both gene content and potential functionality indicates a
392 need for more closed genomes of MTBC sub-lineages and that perhaps a sub-lineage-specific
393 reference genome approach would be appropriate for MTBC WGS. However, although an *ad*
394 *hoc* (e.g. outbreak-specific) reference genome provides a more accurate and precise
395 comparison of strains involved in a specific situation, that reference is only specific to that
396 particular situation/outbreak/population/lineage and cannot be used in another context,
397 making comparisons between lineages and settings difficult.

398 As an alternative to using a single MTBC-wide, lineage or sub-lineage specific reference
399 genome approach, the variable presence of genes within one lineage suggests that an L5
400 pangenome-based reference genome capturing all the known diversity may be required
401 instead. Such a genome would be a composite of all the genes found in all lineages and sub-
402 lineages of the MTBC (i.e. both the core and accessory genome) (McInerney et al. 2020) and
403 is usually represented as a graph instead of a single sequence (Rakocevic et al. 2019; Paten
404 et al. 2017; Martiniano et al. 2019; Marschall et al. 2018; Church et al. 2015). Alternatively, it
405 can be a selection of reference genomes, with mapping to all or a subset undertaken, as has
406 been done with *M. chimaera* and other pathogens (van Ingen et al. 2017; Garimella et al.
407 2019). Other authors have also reported that because of the genetic variability between
408 strains, using a single strain genome as reference genome lacks accuracy (Ballouz, Dobin, and

409 Gillis 2019; Sherman et al. 2019; Li et al. 2010). This MTBC-wide pangenome approach has
410 been suggested before, including for other organisms (Periwal et al. 2015; Meehan et al.
411 2019; Jandrasits et al. 2019; Ballouz, Dobin, and Gillis 2019; Medini et al. 2005; Tettelin et al.
412 2005; Sherman et al. 2019; Li et al. 2010). However, such an approach also has its own
413 drawbacks, including difficulty in mapping reads that bridge the boundaries between
414 accessory genes and the rest of the genome, comparing strains of different lineages including
415 phylogenetic analyses and retention of gene names and codes in clinical use, where H37Rv is
416 deeply embedded (Marschall et al. 2018). The specific sequence of each gene would also need
417 to be chosen for such a reference, with the inferred ancestral genome representative of
418 MTBC lineages approach being the most likely method (Goig et al. 2020, 2018; Comas et al.
419 2013)

420 Another alternative approach is a *de novo* assembly reference free approach (Maretty et al.
421 2017; Iqbal et al. 2012; Meehan et al. 2019; Li et al. 2010) where strains are either assembled
422 into contigs without the use of a reference genome, or compared to each other without first
423 calling SNPs. This would allow for clustering of samples regardless of lineage, e.g. for
424 genotyping or transmission analyses using Mash (software for fast/meta -genome distance
425 estimation technique) (Ondov et al. 2016) but would require new cut-offs for defining clusters
426 and many additional steps for further gene annotation and between sample comparisons,
427 making its clinical use potentially confusing.

428 A limitation of this study is that the completed (PacBio) and short-read (Illumina) genomes
429 were derived from positive cultures, excluding possible minority L5 strain diversity as L5
430 strains are overrepresented in negative cultures (Sanoussi et al. 2017). WGS applied directly
431 on sputum is more and more needed, especially for ancestral lineages (including L5 and L6)

432 where growth in culture is problematic (negative culture or dysgonic isolates that are a
433 challenge to grow for DNA extraction (Sanoussi et al. 2018, 2017)). Also, a limitation regarding
434 the comparison of gene presence/absence in L5 versus L4 (including H37Rv), L6, *M. bovis*
435 complete genomes, is that our study included only a single complete genome of L4, L6 and
436 *M. bovis*, while these lineages may also display variability similar to the intra-L5 variability we
437 observed in our study.

438 In conclusion, the use of a (sub-)lineage reference genome can increase the resolution for
439 strains comparison in comparison to a H37Rv based mapping approach for L5 genome
440 analyses for epidemiology (transmission), phylogeny and sub-lineage determination. Still, the
441 use of a (sub-)lineage reference genome will miss some within-lineage gene differences. For
442 drug-resistance detection in clinical L5 strains or strains of other lineages, H37Rv could still be
443 used as reference genome as resistance-related mutations are usually among core genes
444 (shared across all lineages). The high within-lineage gene content variability suggests the
445 pangenome of MTBC strains may be larger than previously thought, implying a reference-free
446 genome assembly (*de novo assembly*) approach may be needed.

447

448 **Authors and contributors**

449 CNS was involved in conceptualization, methodology, formal analysis, investigation, data
450 curation, writing of the original draft, review/editing the manuscript and visualization.

451 MC was involved in conceptualization, methodology, investigation, data curation, review
452 and editing the manuscript and visualization.

453 BO was involved in methodology, investigation and review/editing the manuscript.

454 IDO was involved in resources and review/editing the manuscript

455 MA was involved in resources.

456 SN was involved in conceptualization, reviewing/editing the manuscript and supervision.

457 JP was involved in investigation, data curation and reviewing/editing the manuscript.

SH was involved in conceptualization, investigation, data curation and reviewing/editing the manuscript.

DY was involved in conceptualization, resources and reviewing/editing the manuscript.

SG was involved in conceptualization and reviewing/editing the manuscript.

LR was involved in conceptualization, resources, review/editing the manuscript and supervision.

DA was involved in resources, review/editing the manuscript and supervision.

BCdJ was involved in conceptualization, methodology, writing of the original draft, review/editing the manuscript and supervision.

CJM was involved in conceptualization, methodology, formal analysis, investigation, writing of the original draft, review/editing the manuscript, supervision and visualization.

Conflict of interest

The authors declare no conflics of interest

Funding information

This work was supported by funds from the Directorate General for Development (DGD), Belgium (FA4 to C.N.S., B.C.D.J., D.A., and L.R.); the European Research Council-INTERRUPTB starting grant (number 311725 to B.C.D.J., C.J.M., and L.R.)

M.C. is supported by ESCMID, Ministerio de Ciencia (RYC-2015-18213 and RTI2018-094399-A-I00) and Generalitat Valenciana (SEJI/2019/011). S.G. is supported by the Swiss National Science Foundation (grants 310030_188888, IZRJZ3_164171, IZLSZ3_170834 and CRSII5_177163) and the European Research Council (883582-ECOEVODRTB).

The funders had no role in study design, data collection and interpretation or the decision to submit the work for publication.

Ethical approval

The PacBio sequenced L5 strain from Benin was isolated during the multicentric OFLOTUB study approved by the National Ethics Committee in Cotonou, Bénin and each of the participating countries (Merle et al. 2012).

Acknowledgements

We would like to acknowledge input from Pim de Rijk Willem and Patrick Beckert for their technical inputs.

492 **Data**

493 Sequence data are available at European Genome- 625 phenome Archive (EGA;
494 https://www.ebi.ac.uk/ega/) under the study accession numbers PRJEB38317 and
495 PRJEB38656. Individual runs accession number for new and published 208 sequences are
496 indicated in Table S7.

497

498 ## Reference

499 Achtman, Mark. 2008. "Evolution, Population Structure, and Phylogeography of Genetically
500 Monomorphic Bacterial Pathogens." *Annual Review of Microbiology* 62 (1): 53–70.
501 https://doi.org/10.1146/annurev.micro.62.081307.162832.

502 Ahmed, F Hafna, A Elaaf Mohamed, Paul D Carr, Brendon M Lee, Karmen Condic-Jurkic, Megan L
503 O'mara, and Colin J Jackson. 2016. "Rv2074 Is a Novel F 420 H 2-Dependent Biliverdin
504 Reductase in Mycobacterium Tuberculosis." https://doi.org/10.1002/pro.2975.

505 Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. "Basic
506 Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10.
507 https://doi.org/10.1016/S0022-2836(05)80360-2.

508 Altschul, Stephen F, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb
509 Miller, and David J Lipman. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein
510 Database Search Programs." *Nucleic Acids Research*. Vol. 25. Oxford University Press.
511 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC146917/pdf/253389.pdf.

512 Ballouz, Sara, Alexander Dobin, and Jesse A. Gillis. 2019. "Is It Time to Change the Reference
513 Genome?" *Genome Biology*. BioMed Central Ltd. https://doi.org/10.1186/s13059-019-1774-4.

514 Behr, M. A., M. A. Wilson, W. P. Gill, H. Salamon, G. K. Schoolnik, S. Rane, and P. M. Small. 1999.
515 "Comparative Genomics of BCG Vaccines by Whole-Genome DNA Microarray." *Science* 284
516 (5419): 1520–23. https://doi.org/10.1126/science.284.5419.1520.

517 Belisle, John T., and Michael G. Sonnenberg. 1998. "Isolation of Genomic DNA from Mycobacteria."
518 In *Mycobacteria Protocols*, 31–44. New Jersey: Humana Press. https://doi.org/10.1385/0-
519 89603-471-2:31.

520 Bentley, Stephen D., Iñaki Comas, Josephine M. Bryant, Danielle Walker, Noel H. Smith, Simon R.
521 Harris, Scott Thurston, et al. 2012. "The Genome of Mycobacterium Africanum West African 2
522 Reveals a Lineage-Specific Locus and Genome Erosion Common to the M. Tuberculosis
523 Complex." Edited by Pamela L. C. Small. *PLoS Neglected Tropical Diseases* 6 (2): e1552.
524 https://doi.org/10.1371/journal.pntd.0001552.

525 Bifani, P, S Moghazeh, B Shopsin, J Driscoll, A Ravikovitch, and B N Kreiswirth. 2000. "Molecular
526 Characterization of Mycobacterium Tuberculosis H37Rv/Ra Variants: Distinguishing the
527 Mycobacterial Laboratory Strain." *Journal of Clinical Microbiology* 38 (9): 3200–3204.
528 https://www.ncbi.nlm.nih.gov/pubmed/10970357.

529 Boritsch, Eva C., Varun Khanna, Alexandre Pawlik, Nadine Honoré, Victor H. Navas, Laurence Ma,
530 Christiane Bouchier, et al. 2016. "Key Experimental Evidence of Chromosomal DNA Transfer
531 among Selected Tuberculosiscausing Mycobacteria." *Proceedings of the National Academy of*
532 *Sciences of the United States of America* 113 (35): 9876–81.
533 https://doi.org/10.1073/pnas.1604921113.

534 Brosch, R, S V Gordon, M Marmiesse, P Brodin, C Buchrieser, K Eiglmeier, T Garnier, et al. 2002. "A
535 New Evolutionary Scenario for the Mycobacterium Tuberculosis Complex." *Proceedings of the*
536 *National Academy of Sciences* 99 (6): 3684–89. https://doi.org/10.1073/pnas.052548299.

537 Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin
538 Bealer, and Thomas L Madden. 2009. "BLAST+: Architecture and Applications." *BMC*
539 *Bioinformatics* 10 (December): 421. https://doi.org/10.1186/1471-2105-10-421.

540 Camus, Jean-Christophe, Melinda J Pryor, Claudine Me, and Stewart T Cole. 2019. "Re-Annotation of
541 the Genome Sequence of Mycobacterium Tuberculosis H37Rv." Vol. 24. www.sanger.ac.uk.

542 Cavet, Jennifer S., Alison I. Graham, Wenmao Meng, and Nigel J. Robinson. 2003. "A Cadmium-Lead-
543 Sensing ArsR-SmtB Repressor with Novel Sensory Sites. Complementary Metal Discrimination
544 by NmtR and CmtR in a Common Cytosol." *Journal of Biological Chemistry* 278 (45): 44560–66.
545 https://doi.org/10.1074/jbc.M307877200.

546 Church, Deanna M., Valerie A. Schneider, Karyn Meltz Steinberg, Michael C. Schatz, Aaron R.
547 Quinlan, Chen Shan Chin, Paul A. Kitts, et al. 2015. "Extending Reference Assembly Models."
548 *Genome Biology* 16 (1): 1–5. https://doi.org/10.1186/s13059-015-0587-3.

549 Cole, S. T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, et al. 1998.
550 "Deciphering the Biology of Mycobacterium Tuberculosis from the Complete Genome
551 Sequence." *Nature* 393 (6685): 537–44. https://doi.org/10.1038/31159.

552 Coll, Francesc, Ruth McNerney, José Afonso Guerra-Assunção, Judith R. Glynn, João Perdigão, Miguel
553 Viveiros, Isabel Portugal, Arnab Pain, Nigel Martin, and Taane G. Clark. 2014. "A Robust SNP
554 Barcode for Typing Mycobacterium Tuberculosis Complex Strains." *Nature Communications* 5
555 (1): 4812. https://doi.org/10.1038/ncomms5812.

556 Comas, Iñaki, Mireia Coscolla, Tao Luo, Sonia Borrell, Kathryn E Holt, Midori Kato-Maeda, Julian
557 Parkhill, et al. 2013. "Out-of-Africa Migration and Neolithic Coexpansion of Mycobacterium
558 Tuberculosis with Modern Humans." *Nature Genetics* 45 (10). https://doi.org/10.1038/ng.2744.

559 Coscolla, M, Daniela Brites, Fabrizio Menardo, Chloe Loiseau, Isaac Darko Otchere, Adwoa Asante-
560 Poku, Prince Asare, et al. 2020. "Phylogenomics of Mycobacterium Africanum Reveals a New
561 Lineage and a Complex Evolutionary History." *BioRxiv* 17 (June): 19.
562 https://doi.org/10.1101/2020.06.10.141788.

563 Coscolla, Mireia, and Sebastien Gagneux. 2014. "Consequences of Genomic Diversity in
564 Mycobacterium Tuberculosis." *Seminars in Immunology* 26 (6): 431–44.
565 https://doi.org/10.1016/j.smim.2014.09.012.

566 Darko Otchere, Isaac, Mireia Coscollá, Leonor Sánchez-Busó, Adwoa Asante-Poku, Daniela Brites,
567 Chloe Loiseau, Conor Meehan, et al. 2018. "Comparative Genomics of Mycobacterium
568 Africanum Lineage 5 and Lineage 6 from Ghana Suggests Distinct Ecological Niches." *SCIENTIfIC*
569 *RepoRtS* | 8: 11269. https://doi.org/10.1038/s41598-018-29620-2.

570 Darling, Aaron C E, Bob Mau, Frederick R. Blattner, and Nicole T. Perna. 2004. "Mauve: Multiple
571 Alignment of Conserved Genomic Sequence with Rearrangements." *Genome Research* 14 (7):
572 1394–1403. https://doi.org/10.1101/gr.2289704.

573 DeJong, Bouke C, Philip C Hill, Alex Aiken, Timothy Awine, Martin Antonio, Ifedayo M Adetifa, Dolly J
574 Jackson-Sillah, et al. 2008. "Progression to Active Tuberculosis, but Not Transmission, Varies by
575 M. Tuberculosis Lineage in The Gambia." *J Infect Dis* 198 (7): 1037–43.
576 https://doi.org/10.1086/591504.

577 Dick, Thomas, Ujjini Manjunatha, Barbara Kappes, and Martin Gengenbacher. 2010. "Vitamin B6

578 Biosynthesis Is Essential for Survival and Virulence of Mycobacterium Tuberculosis." *Molecular*
579 *Microbiology* 78 (4): 980–88. https://doi.org/10.1111/j.1365-2958.2010.07381.x.

580 Firdessa, Rebuma, Stefan Berg, Elena Hailu, Esther Schelling, Balako Gumi, Girume Erenso,
581 Endalamaw Gadisa, et al. 2013. "Mycobacterial Lineages Causing Pulmonary and
582 Extrapulmonary Tuberculosis, Ethiopia." *Emerging Infectious Diseases* 19 (3): 460–63.
583 https://doi.org/10.3201/eid1903.120256.

584 Gagneux, Sebastien. 2018. "Ecology and Evolution of Mycobacterium Tuberculosis." *Nature Reviews*
585 *Microbiology*. https://doi.org/10.1038/nrmicro.2018.8.

586 Gagneux, Sebastien, Kathryn Deriemer, Tran Van, Midori Kato-Maeda, Bouke C De Jong, Sujatha
587 Narayanan, Mark Nicol, et al. 2006. "Variable Host-Pathogen Compatibility in Mycobacterium
588 Tuberculosis." www.pnas.orgcgidoi10.1073pnas.0511240103.

589 Gao, Juehua, Jie Liao, and Guang-Yu Yang. 2009a. "Review Article CAAX-Box Protein, Prenylation
590 Process and Carcinogenesis." *Am J Transl Res*. Vol. 1. www.ajtr.org/AJTR905003.

591 Gao, Juehua, Jie Liao, and Guang Yu Yang. 2009b. "CAAX-Box Protein, Prenylation Process and
592 Carcinogenesis." *American Journal of Translational Research*.

593 Garimella, Kiran V, Zamin Iqbal, Michael A. Krause, Susana Campino, Mihir Kekre, Eleanor Drury,
594 Dominic Kwiatkowski, Juliana M. Sa, Thomas E. Wellems, and Gil McVean. 2019. "Detection of
595 Simple and Complex de Novo Mutations without, with, or with Multiple Reference Sequences."
596 *BioRxiv*, July, 698910. https://doi.org/10.1101/698910.

597 Gehre, Florian, Jacob Otu, Kathryn DeRiemer, Paola Florez de Sessions, Martin L. Hibberd, Wim
598 Mulders, Tumani Corrah, Bouke C. de Jong, and Martin Antonio. 2013. "Deciphering the
599 Growth Behaviour of Mycobacterium Africanum." *PLoS Neglected Tropical Diseases* 7 (5).
600 https://doi.org/10.1371/journal.pntd.0002220.

601 Ghodousi, Arash, Alamdar Hussain Rizvi, Aurangzaib Quadir Baloch, Abdul Ghafoor, Faisal Masood,
602 Mehmood Qadir, Emanuele Borroni, Alberto Trovato, Sabira Tahseen, and Daniela Maria. 2019.
603 "Acquisition of Cross-Resistance to Bedaquiline and Clofazimine Following Treatment for
604 Tuberculosis in 2 Pakistan." https://doi.org/10.1101/630012.

605 Goig, Galo A., Silvia Blanco, Alberto L. Garcia-Basteiro, and Iñaki Comas. 2020. "Contaminant DNA in
606 Bacterial Sequencing Experiments Is a Major Source of False Genetic Variability." *BMC Biology*
607 18 (1): 24. https://doi.org/10.1186/s12915-020-0748-z.

608 Goig, Galo A, Silvia Blanco, Alberto L. Garcia-Basteiro, and Iñaki Comas. 2019. "Contaminant DNA in
609 Bacterial Sequencing Experiments Is a Major Source of False Genetic Variability." *BioRxiv*.

610 Goig, Galo A, Silvia Blanco, Alberto L Garcia-Basteiro, and Iñaki Comas. 2018. "Pervasive
611 Contaminations in Sequencing Experiments Are a Major Source of False Genetic Variability: A
612 Mycobacterium Tuberculosis Meta-Analysis." https://doi.org/10.1101/403824.

613 Gómez-Rubio, Virgilio. 2017. "Ggplot2 - Elegant Graphics for Data Analysis (2nd Edition) ." *Journal of*
614 *Statistical Software* 77 (Book Review 2): 3–5. https://doi.org/10.18637/jss.v077.b02.

615 Gonzalo-Asensio, Jesús, Wladimir Malaga, Alexandre Pawlik, Catherine Astarie-Dequeker, Charlotte
616 Passemar, Flavie Moreau, Françoise Laval, et al. 2014. "Evolutionary History of Tuberculosis
617 Shaped by Conserved Mutations in the PhoPR Virulence Regulator." *Proceedings of the*
618 *National Academy of Sciences of the United States of America* 111 (31): 11491–96.
619 https://doi.org/10.1073/pnas.1406693111.

620 Gordon, Stephen V., Roland Brosch, Alain Billault, Thierry Garnier, Karin Eiglmeier, and Stewart T.

621 Cole. 1999. “Identification of Variable Regions in the Genomes of Tubercle Bacilli Using
622 Bacterial Artificial Chromosome Arrays.” *Molecular Microbiology* 32 (3): 643–55.
623 https://doi.org/10.1046/j.1365-2958.1999.01383.x.

624 Grinter, Rhys, Blair Ney, Rajini Brammananth, Christopher K. Barlow, Paul R.F. Cordero, David L.
625 Gillett, Thierry Izoré, et al. 2020. “Cellular and Structural Basis of Synthesis of the Unique
626 Intermediate Dehydro-F420-0 in Mycobacteria.” *BioRxiv*, February, 2020.02.27.968891.
627 https://doi.org/10.1101/2020.02.27.968891.

628 Gu, J, H Li, M Li, C Vuong, M Otto, Y Wen, and Q Gao. 2005. “Bacterial Insertion Sequence IS256 as a
629 Potential Molecular Marker to Discriminate Invasive Strains from Commensal Strains of
630 Staphylococcus Epidermidis.” *The Journal of Hospital Infection* 61 (4): 342–48.
631 https://doi.org/10.1016/j.jhin.2005.04.017.

632 Herrmann, Jean Louis, Robin Delahay, Alex Gallagher, Brian Robertson, and Douglas Young. 2000.
633 “Analysis of Post-Translational Modification of Mycobacterial Proteins Using a Cassette
634 Expression System.” *FEBS Letters* 473 (3): 358–62. https://doi.org/10.1016/S0014-
635 5793(00)01553-2.

636 Ingen, Jakko van, Thomas A. Kohl, Katharina Kranzer, Barbara Hasse, Peter M. Keller, Anna Katarzyna
637 Szafrańska, Doris Hillemann, et al. 2017. “Global Outbreak of Severe Mycobacterium Chimaera
638 Disease after Cardiac Surgery: A Molecular Epidemiological Study.” *The Lancet Infectious*
639 *Diseases* 17 (10): 1033–41. https://doi.org/10.1016/S1473-3099(17)30324-9.

640 Intemann, Christopher D., Thorsten Thye, Stefan Niemann, Edmund N. L. Browne, Margaret Amanua
641 Chinbuah, Anthony Enimil, John Gyapong, et al. 2009. “Autophagy Gene Variant IRGM −261T
642 Contributes to Protection from Tuberculosis Caused by Mycobacterium Tuberculosis but Not by
643 M. Africanum Strains.” Edited by William Bishai. *PLoS Pathogens* 5 (9): e1000577.
644 https://doi.org/10.1371/journal.ppat.1000577.

645 Iqbal, Zamin, Mario Caccamo, Isaac Turner, Paul Flicek, and Gil Mcvean. 2012. “De Novo Assembly
646 and Genotyping of Variants Using Colored de Bruijn Graphs.” https://doi.org/10.1038/ng.1028.

647 Jandrasits, Christine, Stefan Kröger, Walter Haas, and Bernhard Y. Renard. 2019. “Computational
648 Pan-Genome Mapping and Pairwise SNP-Distance Improve Detection of Mycobacterium
649 Tuberculosis Transmission Clusters.” Edited by Sergei L. Kosakovsky Pond. *PLOS Computational*
650 *Biology* 15 (12): e1007527. https://doi.org/10.1371/journal.pcbi.1007527.

651 Jong, Bouke C De, Martin Antonio, and Sebastien Gagneux. n.d. “Mycobacterium Africanum-Review
652 of an Important Cause of Human Tuberculosis in West Africa.” Accessed May 21, 2019.
653 https://doi.org/10.1371/journal.pntd.0000744.

654 Kapopoulou, Adamandia, Jocelyne M. Lew, and Stewart T. Cole. 2011. “The MycoBrowser Portal: A
655 Comprehensive and Manually Annotated Resource for Mycobacterial Genomes.” *Tuberculosis*
656 91 (1): 8–13. https://doi.org/10.1016/j.tube.2010.09.006.

657 Kato-Maeda, Midori, Jeanne T Rhee, Thomas R Gingeras, Hugh Salamon, Jorg Drenkow, Nat
658 Smittipat, and Peter M Small. 2001. “Comparing Genomes within the Species Mycobacterium
659 Tuberculosis.” https://doi.org/10.1101/gr166401.

660 Kavanagh, K L, H Jçrnvall, B Persson, and U Oppermann. 2008. “The SDR Superfamily: Functional and
661 Structural Diversity within a Family of Metabolic and Regulatory Enzymes” 171: 77.
662 https://doi.org/10.1007/s00018-008-8588-y.

663 Kohl, Thomas Andreas, Christian Utpatel, Viola Schleusener, Maria Rosaria De Filippo, Patrick
664 Beckert, Daniela Maria Cirillo, and Stefan Niemann. 2018. “MTBseq: A Comprehensive Pipeline

for Whole Genome Sequence Analysis of Mycobacterium Tuberculosis Complex Isolates." https://doi.org/10.7717/peerj.5895.

Leao, Sylvia, Anandi Martin, Gloria Isabel Mejia, Francoise Portaels, and Por- Taels F Leao SC, Martin A, Mejia GI, Palomino JC, Robledo J, Telles MAS. 2004. "Practical Handbook for the Phenotypic and Genotypic Identification of Mycobacteria.," 77–126. https://doi.org/10.1007/BF01045823.

Lee, Robyn S., Jean François Proulx, Fiona McIntosh, Marcel A. Behr, and William P. Hanage. 2020. "Previously Undetected Super-Spreading of Mycobacterium Tuberculosis Revealed by Deep Sequencing." *ELife* 9 (February). https://doi.org/10.7554/eLife.53245.

Lee, Robyn S, and Marcel A Behr. 2016. "Does Choice Matter? Reference-Based Alignment for Molecular Epidemiology of Tuberculosis." *Journal of Clinical Microbiology* 54 (7): 1891–95. https://doi.org/10.1128/JCM.00364-16.

Lew, Jocelyne M., Adamandia Kapopoulou, Louis M. Jones, and Stewart T. Cole. 2011a. "TubercuList - 10 Years After." *Tuberculosis* 91 (1): 1–7. https://doi.org/10.1016/j.tube.2010.09.008.

———. 2011b. "TubercuList - 10 Years After." *Tuberculosis* 91 (1): 1–7. https://doi.org/10.1016/j.tube.2010.09.008.

Li, Ruiqiang, Yingrui Li, Hancheng Zheng, Ruibang Luo, Hongmei Zhu, Qibin Li, Wubin Qian, et al. 2010. "Building the Sequence Map of the Human Pan-Genome." *Nature Biotechnology* 28 (1): 57–62. https://doi.org/10.1038/nbt.1596.

Magnus, K. 1966. "Epidemiological Basis of Tuberculosis Eradication 3. Risk of Pulmonary Tuberculosis after Human and Bovine Infection *." *Bull. Org. Mond. Sante*. Vol. 35. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2476032/pdf/bullwho00607-0027.pdf.

Malone, Kerri M, Damien Farrell, Tod P Stuber, Olga T Schubert, Ruedi Aebersold, Suelee Robbe-Austerman, and Stephen V Gordon. 2017. "Updated Reference Genome Sequence and Annotation of Mycobacterium Bovis AF2122/97." *Genome Announcements* 5 (14): e00157-17. https://doi.org/10.1128/genomeA.00157-17.

Marcus, Sarah A, Sarah W Sidiropoulos, Howard Steinberg, and Adel M Talaat. 2016. "CsoR Is Essential for Maintaining Copper Homeostasis in Mycobacterium Tuberculosis." https://doi.org/10.1371/journal.pone.0151816.

Maretty, Lasse, Jacob Malte Jensen, Bent Petersen, Jonas andreas sibbesen, siyang Liu, Palle Villesen, Laurits skov, et al. 2017. "Sequencing and de Novo Assembly of 150 Genomes from Denmark as a Population Reference." *Nature Publishing Group* 548. https://doi.org/10.1038/nature23264.

Marschall, Tobias, Manja Marz, Thomas Abeel, Louis Dijkstra, Bas E. Dutilh, Ali Ghaffaari, Paul Kersey, et al. 2018. "Computational Pan-Genomics: Status, Promises and Challenges." *Briefings in Bioinformatics* 19 (1): 118–35. https://doi.org/10.1093/bib/bbw089.

Martiniano, Rui, Erik Garrison, Eppie R Jones, Andrea Manica, and Richard Durbin. 2019. "Removing Reference Bias in Ancient DNA Data Analysis by Mapping to a Sequence Variation Graph." *BioRxiv*, September, 782755. https://doi.org/10.1101/782755.

McInerney, James O., Fiona J. Whelan, Maria Rosa Domingo-Sananes, Alan McNally, and Mary J. O'Connell. 2020. "Pangenomes and Selection: The Public Goods Hypothesis." In *The Pangenome*, 151–67. Springer International Publishing. https://doi.org/10.1007/978-3-030-38281-0_7.

Medini, Duccio, Claudio Donati, Hervé Tettelin, Vega Masignani, and Rino Rappuoli. 2005. "The Microbial Pan-Genome." *Current Opinion in Genetics and Development*. Elsevier Current

Trends. https://doi.org/10.1016/j.gde.2005.09.006.

Meehan, Conor J., Pieter Moris, Thomas A. Kohl, Jūlija Pečerska, Suriya Akter, Matthias Merker, Christian Utpatel, et al. 2018. "The Relationship between Transmission Time and Clustering Methods in Mycobacterium Tuberculosis Epidemiology." *EBioMedicine* 37 (November): 410–16. https://doi.org/10.1016/j.ebiom.2018.10.013.

Meehan, Conor J, Galo A Goig, Thomas A Kohl, Lennert Verboven, Anzaan Dippenaar, Matthew Ezewudo, Maha R Farhat, et al. 2019. "Whole Genome Sequencing of Mycobacterium Tuberculosis: Current Standards and Open Issues." *Nature Reviews Microbiology*. https://doi.org/10.1038/s41579-019-0214-5.

Merker, Matthias, Thomas A Kohl, Ivan Barilar, Sönke Andres, Philip W Fowler, Erja Chryssanthou, Kristian Ängeby, et al. 2020. "Phylogenetically Informative Mutations in Genes Implicated in Antibiotic Resistance in Mycobacterium Tuberculosis Complex." *Genome Medicine*. https://doi.org/10.1186/s13073-020-00726-5.

Merle, Corinne S.C., Charalambos Sismanidis, Oumou B. Sow, Martin Gninafon, John Horton, Olivier Lapujade, Mame B. Lo, et al. 2012. "A Pivotal Registration Phase III, Multicenter, Randomized Tuberculosis Controlled Trial: Design Issues and Lessons Learnt from the Gatifloxacin for TB (OFLOTUB) Project." *Trials* 13 (May). https://doi.org/10.1186/1745-6215-13-61.

Murugesan, Saravanan, Stalin Mani, Indhumathy Kuppusamy, and Padma Krishnan. 2018. "Role of Insertion Sequence Element Is256 as a Virulence Marker and Its Association with Biofilm Formation among Methicillin-Resistant Staphylococcus Epidermidis from Hospital and Community Settings in Chennai, South India." *Indian Journal of Medical Microbiology* 36 (1): 124. https://doi.org/10.4103/ijmm.IJMM_17_276.

Nebenzahl-Guimaraes, Hanna, Solomon A. Yimer, Carol Holm-Hansen, Jessica De Beer, Roland Brosch, and Dick Van Soolingen. 2016. "Genomic Characterization of Mycobacterium Tuberculosis Lineage 7 and a Proposed Name: 'Aethiops Vetus.'" *Microbial Genomics* 2 (6). https://doi.org/10.1099/mgen.0.000063.

Ngabonziza, Jean Claude Semuto, Chloé Loiseau, Michael Marceau, Agathe Jouet, Fabrizio Menardo, Oren Tzfadia, Esdras Belamo Niyigena, et al. 2020. *A Sister Lineage of the Mycobacterium Tuberculosis Complex Discovered in the African Great Lakes Region*. *BioRxiv*. https://doi.org/10.1101/2020.01.20.912998.

Niemann, S., T. Kubica, F. C. Bange, O. Adjei, E. N. Browne, M. A. Chinbuah, R. Diel, et al. 2004. "The Species Mycobacterium Africanum in the Light of New Molecular Markers." *Journal of Clinical Microbiology* 42 (9): 3958–62. https://doi.org/10.1128/JCM.42.9.3958-3962.2004.

Norman, Anders, Dorte Bek Folkvardsen, Søren Overballe-Petersen, and Troels Lillebaek. 2019. "Complete Genome Sequence of Mycobacterium Tuberculosis DKC2, the Predominant Danish Outbreak Strain." https://doi.org/10.1128/MRA.

O'Toole, Ronan F., and Sanjay S. Gautam. 2017. "Limitations of the Mycobacterium Tuberculosis Reference Genome H37Rv in the Detection of Virulence-Related Loci." *Genomics* 109 (5–6): 471–74. https://doi.org/10.1016/j.ygeno.2017.07.004.

Ofori-Anyinam, Boatema, Fatoumatta Kanuteh, Schadrac C. Agbla, Ifedayo Adetifa, Catherine Okoi, Gregory Dolganov, Gary Schoolnik, et al. 2016. "Impact of the Mycobaterium Africanum West Africa 2 Lineage on TB Diagnostics in West Africa: Decreased Sensitivity of Rapid Identification Tests in The Gambia." *PLoS Neglected Tropical Diseases* 10: 1–12. https://doi.org/10.1371/journal.pntd.0004801.

Ofori-Anyinam, Boatema, Abi Janet Riley, Tijan Jobarteh, Ensa Gitteh, Binta Sarr, Tutty Isatou Faal-Jawara, Leen Rigouts, et al. 2020. "Comparative Genomics Shows Differences in the Electron Transport and Carbon Metabolic Pathways of Mycobacterium Africanum Relative to Mycobacterium Tuberculosis and Suggests an Adaptation to Low Oxygen Tension." *Tuberculosis* 120 (January): 101899. https://doi.org/10.1016/j.tube.2020.101899.

Ondov, Brian D., Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. 2016. "Mash: Fast Genome and Metagenome Distance Estimation Using MinHash." *Genome Biology* 17 (1): 132. https://doi.org/10.1186/s13059-016-0997-x.

Paten, Benedict, Adam M. Novak, Jordan M. Eizenga, and Erik Garrison. 2017. "Genome Graphs and the Evolution of Genome Inference." *Genome Research*. Cold Spring Harbor Laboratory Press. https://doi.org/10.1101/gr.214155.116.

Periwal, Vinita, Ashok Patowary, Shamsudheen Karuthedath Vellarikkal, Anju Gupta, Meghna Singh, Ashish Mittal, Shamini Jeyapaul, et al. 2015. "Comparative Whole-Genome Analysis of Clinical Isolates Reveals Characteristic Architecture of Mycobacterium Tuberculosis Pangenome." https://doi.org/10.1371/journal.pone.0122979.

Phelan, Jody, Paola Florez De Sessions, Leopold Tientcheu, Joao Perdigao, Diana Machado, Rumina Hasan, Zahra Hasan, et al. 2018. "Methylation in Mycobacterium Tuberculosis Is Lineage Specific with Associated Mutations Present Globally." *Scientific Reports* 8 (1): 1–7. https://doi.org/10.1038/s41598-017-18188-y.

R Core Team, R. 2019. "R: A Language and Environment for Statistical Computing." Edited by R Development Core Team. *R Foundation for Statistical Computing*. R Foundation for Statistical Computing. R Foundation for Statistical Computing. https://doi.org/10.1007/978-3-540-74686-7.

Rademacher, Corinna, and Bernd Masepohl. 2012. "Copper-Responsive Gene Regulation in Bacteria." https://doi.org/10.1099/mic.0.058487-0.

Rakocevic, Goran, Vladimir Semenyuk, Wan Ping Lee, James Spencer, John Browning, Ivan J. Johnson, Vladan Arsenijevic, et al. 2019. "Fast and Accurate Genomic Analyses Using Genome Graphs." *Nature Genetics* 51 (2): 354–62. https://doi.org/10.1038/s41588-018-0316-4.

Reiling, Norbert, Susanne Homolka, Kerstin Walter, Julius Brandenburg, Lisa Niwinski, Martin Ernst, Christian Herzmann, et al. 2013. "Clade-Specific Virulence Patterns of Mycobacterium Tuberculosis Complex Strains in Human Primary Macrophages and Aerogenically Infected Mice." *MBio* 4 (4). https://doi.org/10.1128/mBio.00250-13.

Rengarajan, Jyothi, Barry R Bloom, and Eric J Rubin. 2005. "Genome-Wide Requirements for Mycobacterium Tuberculosis Adaptation and Survival in Macrophages." *PNAS June*. Vol. 7. www.pnas.orgcgidoi10.1073pnas.0503272102.

Riojas, Marco A, Katya J Mcgough, Cristin J Rider-Riojas, Nalin Rastogi, and Manzour Hernando Hazbón. 2018. "Phylogenomic Analysis of the Species of the Mycobacterium Tuberculosis Complex Demonstrates That Mycobacterium Africanum, Mycobacterium Bovis, Mycobacterium Caprae, Mycobacterium Microti and Mycobacterium Pinnipedii Are Later Heterotypic Synonyms of Mycob." https://doi.org/10.1099/ijsem.0.002507.

Rowland, Jennifer L, and Michael Niederweis. 2012. "Resistance Mechanisms of Mycobacterium Tuberculosis against Phagosomal Copper Overload." https://doi.org/10.1016/j.tube.2011.12.006.

S.R. Pattyn, F. Portaels, L. Spanoghe, and J. Magos. 1970. “Further Studies on African Strains of Mycobacterium Tuberculosis. Comparison with M. Bovis and M. Microti.” Ann. Soc. belge Méd. Trop. https://www.ncbi.nlm.nih.gov/pubmed/4996738.

Samanovic, Marie I., Chen Ding, Dennis J. Thiele, and K. Heran Darwin. 2012. “Copper in Microbial Pathogenesis: Meddling with the Metal.” *Cell Host & Microbe* 11 (2): 106–15. https://doi.org/10.1016/j.chom.2012.01.009.

Sanoussi, C. N’Dira, Dissou Affolabi, Leen Rigouts, Séverin Anagonou, and Bouke de Jong. 2017. “Genotypic Characterization Directly Applied to Sputum Improves the Detection of Mycobacterium Africanum West African 1, under-Represented in Positive Cultures.” *PLOS Neglected Tropical Diseases* 11 (9): e0005900. https://doi.org/10.1371/journal.pntd.0005900.

Sanoussi, C. N’Dira, Bouke C De Jong, Mathieu Odoun, Karamatou Arekpa, Moulikatou Ali Ligali, Ousman Bodi, Simon Harris, et al. 2018. “Low Sensitivity of the MPT64 Identification Test to Detect Lineage 5 of the Mycobacterium Tuberculosis Complex.” https://doi.org/10.1099/jmm.0.000846.

Seemann, T. 2014. “Prokka: Rapid Prokaryotic Genome Annotation.” *Bioinformatics* 30 (14): 2068–69. https://doi.org/10.1093/bioinformatics/btu153.

Selengut, Jeremy D, and Daniel H Haft. 2010. “Unexpected Abundance of Coenzyme F 420-Dependent Enzymes in Mycobacterium Tuberculosis and Other Actinobacteria †.” *JOURNAL OF BACTERIOLOGY* 192 (21): 5788–98. https://doi.org/10.1128/JB.00425-10.

Sellés Vidal, Lara, Ciarán L. Kelly, Paweł M. Mordaka, and John T. Heap. 2018. “Review of NAD(P)H-Dependent Oxidoreductases: Properties, Engineering and Application.” *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1866 (2): 327–47. https://doi.org/10.1016/j.bbapap.2017.11.005.

Sherman, Rachel M., Juliet Forman, Valentin Antonescu, Daniela Puiu, Michelle Daya, Nicholas Rafaels, Meher Preethi Boorgula, et al. 2019. “Assembly of a Pan-Genome from Deep Sequencing of 910 Humans of African Descent.” *Nature Genetics*. Nature Publishing Group. https://doi.org/10.1038/s41588-018-0273-y.

Sreevatsan, Srinand, Xi Pan, Kathryn E. Stockbauer, Nancy D. Connell, Barry N. Kreiswirth, Thomas S. Whittam, and James M. Musser. 1997. “Restricted Structural Gene Polymorphism in the Mycobacterium Tuberculosis Complex Indicates Evolutionarily Recent Global Dissemination.” *Proceedings of the National Academy of Sciences of the United States of America* 94 (18): 9869–74. https://doi.org/10.1073/pnas.94.18.9869.

Szklarczyk, Damian, John H. Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, et al. 2017. “The STRING Database in 2017: Quality-Controlled Protein-Protein Association Networks, Made Broadly Accessible.” *Nucleic Acids Research* 45 (D1): D362–68. https://doi.org/10.1093/nar/gkw937.

Tatusov, Roman L, Michael Y Galperin, Darren A Natale, and Eugene V Koonin. 2000. “The COG Database: A Tool for Genome-Scale Analysis of Protein Functions and Evolution.” *Nucleic Acids Research*. Vol. 28. http://www.ncbi.nlm.nih.gov/COG.

Tettelin, Hervé, Vega Masignani, Michael J. Cieslewicz, Claudio Donati, Duccio Medini, Naomi L. Ward, Samuel V. Angiuoli, et al. 2005. “Genome Analysis of Multiple Pathogenic Isolates of Streptococcus Agalactiae: Implications for the Microbial ‘Pan-Genome.’” *Proceedings of the National Academy of Sciences of the United States of America* 102 (39): 13950–55. https://doi.org/10.1073/pnas.0506758102.

840 Thye, Thorsten, Stefan Niemann, Kerstin Walter, Susanne Homolka, Christopher D. Intemann,
841 Margaret Amanua Chinbuah, Anthony Enimil, et al. 2011. "Variant G57E of Mannose Binding
842 Lectin Associated with Protection against Tuberculosis Caused by Mycobacterium Africanum
843 but Not by M. Tuberculosis." *PLoS ONE* 6 (6). https://doi.org/10.1371/journal.pone.0020908.

844 Tsolaki, Anthony G., Sebastien Gagneux, Alexander S. Pym, Yves Olivier L. Goguet De La Salmoniere,
845 Barry N. Kreiswirth, Dick Van Soolingen, and Peter M. Small. 2005. "Genomic Deletions Classify
846 the Beijing/W Strains as a Distinct Genetic Lineage of Mycobacterium Tuberculosis." *Journal of*
847 *Clinical Microbiology* 43 (7): 3185–91. https://doi.org/10.1128/JCM.43.7.3185-3191.2005.

848 Walker, Timothy M, Camilla LC Ip, Ruth H Harrell, Jason T Evans, Georgia Kapatai, Martin J Dedicoat,
849 David W Eyre, et al. 2013. "Whole-Genome Sequencing to Delineate Mycobacterium
850 Tuberculosis Outbreaks: A Retrospective Observational Study." *The Lancet Infectious Diseases*
851 13 (2): 137–46. https://doi.org/10.1016/S1473-3099(12)70277-3.

852 Ward, Sarah K, Elizabeth A Hoye, and Adel M Talaat. 2008. "The Global Responses of Mycobacterium
853 Tuberculosis to Physiological Levels of Copper †." *JOURNAL OF BACTERIOLOGY* 190 (8): 2939–
854 46. https://doi.org/10.1128/JB.01847-07.

855 Warnes, Gregory R., Ben Bolker, Gregor Gorjanc, Gabor Grothendieck, Ales Korosec, Thomas Lumley,
856 Don MacQueen, Arni Magnusson, Jim Rogers, and and others. 2017. "CRAN - Package Gdata."
857 https://mran.microsoft.com/snapshot/2017-12-11/web/packages/gdata/index.html.

858 WHO. 2016. "THE SHORTER MDR-TB REGIMEN FEATURES OF THE SHORTER MDR-TB REGIMEN
859 REGIMEN COMPOSITION 4-6 Km-Mfx-Pto-Cfz-Z-H High-Dose-E / 5 Mfx-Cfz-Z-E."
860 www.who.int/tb.

861 ———. 2019. *WHO Consolidated Guidelines on Drug-Resistant Tuberculosis Treatment*.
862 http://apps.who.int/bookorders.

863 Zhang, Shuo, Jiazhen Chen, Peng Cui, Wanliang Shi, Wenhong Zhang, and Ying Zhang. 2015.
864 "Identification of Novel Mutations Associated with Clofazimine Resistance in Mycobacterium
865 Tuberculosis." https://doi.org/10.1093/jac/dkv150.

866

867

868

869

870

871

872

873

874

875

876

877

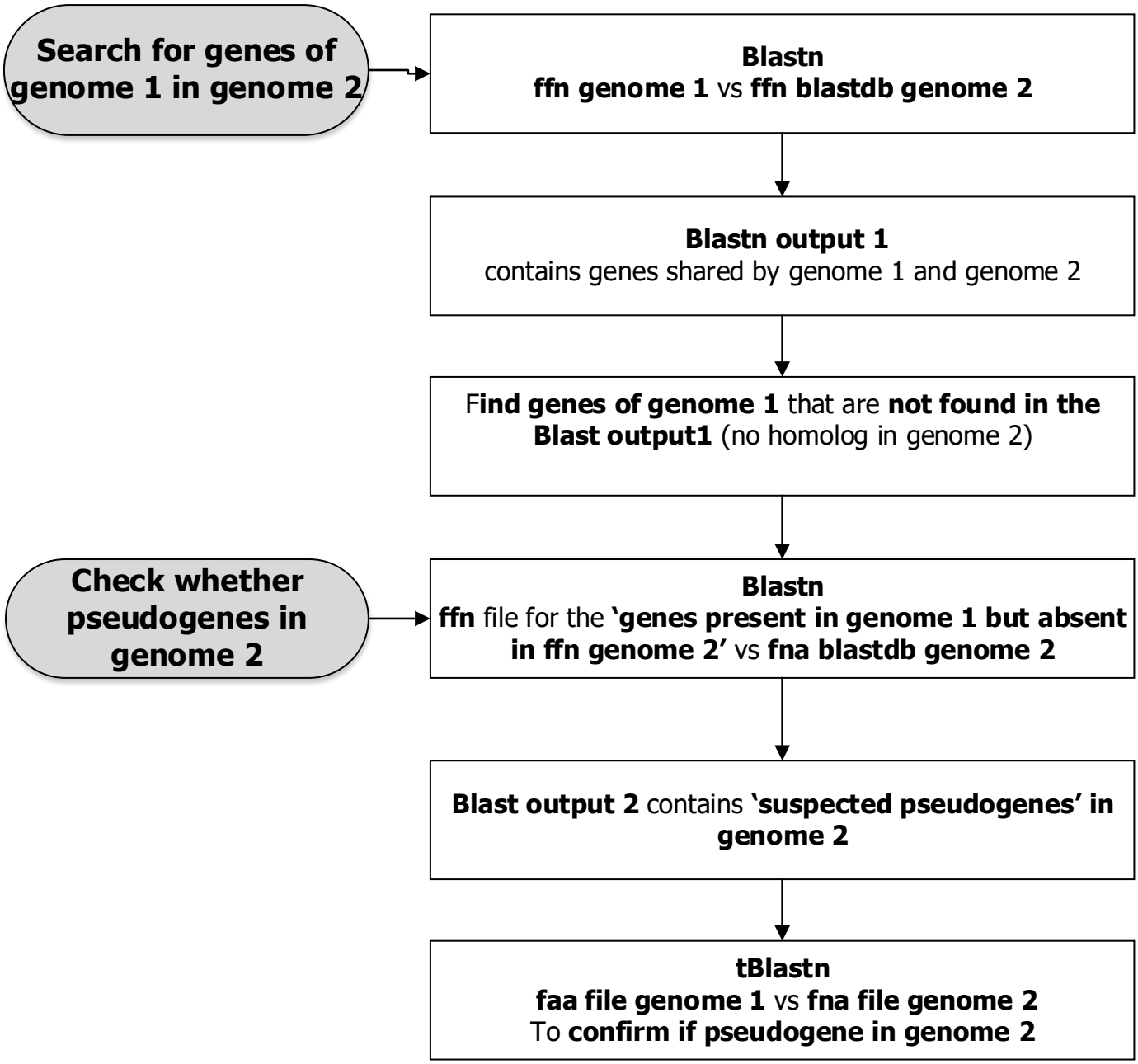
878

879

880

881

882



883

884 **Figure 1. Methodology for finding gene presence-absence between two genomes**

885

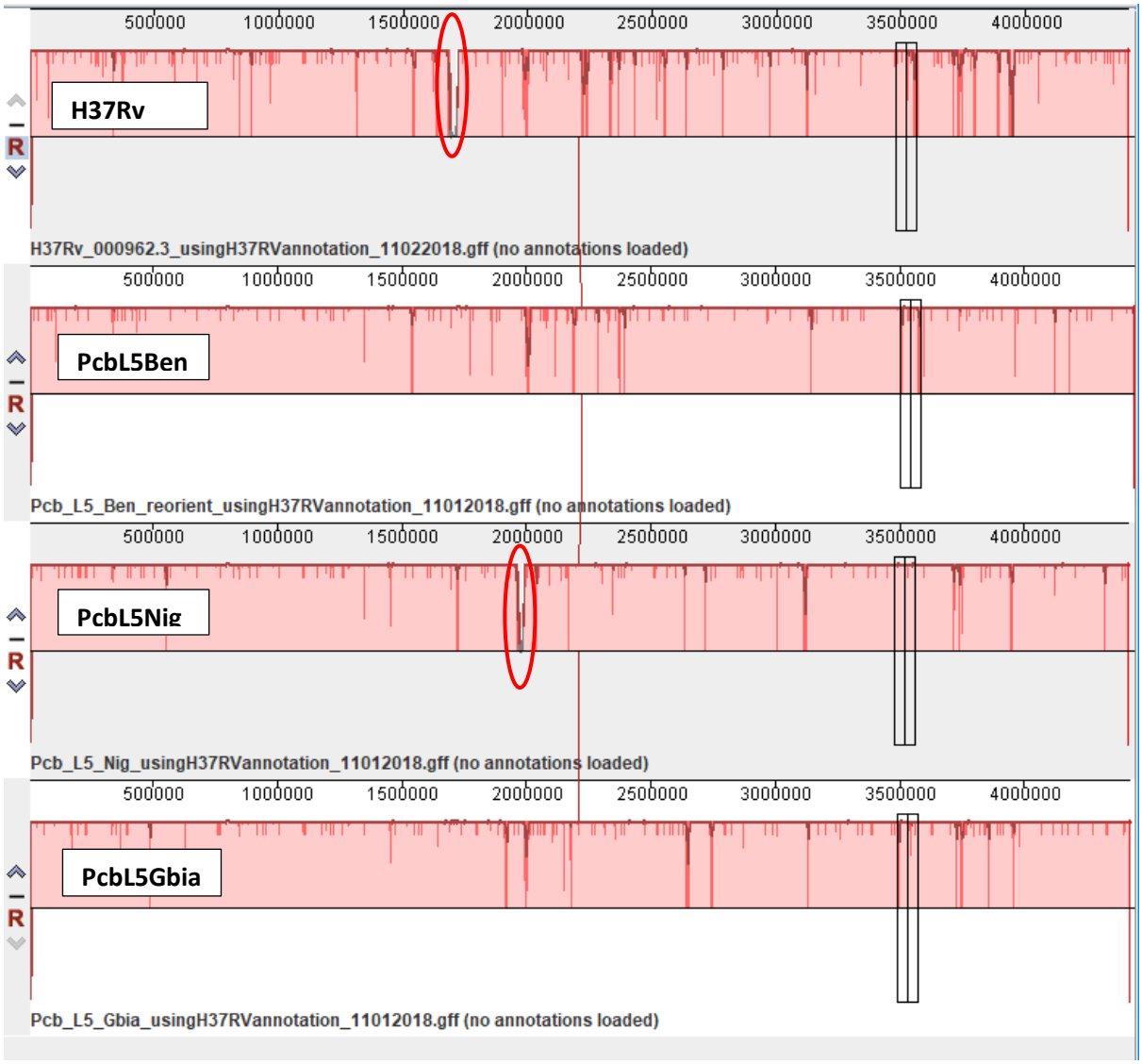


Figure 2. Visualization of possible rearrangements using Mauve pipeline: H37Rv (L4) genome and three completed PacBio genomes of *M. africanum* L5 isolates from Benin, The Gambia and Nigeria.

The region circled in red (one on the H37Rv genome, and one on the Nigerian completed PacBio genome) represents the regions absent in that strain genome but present in the other three strains complete genomes.

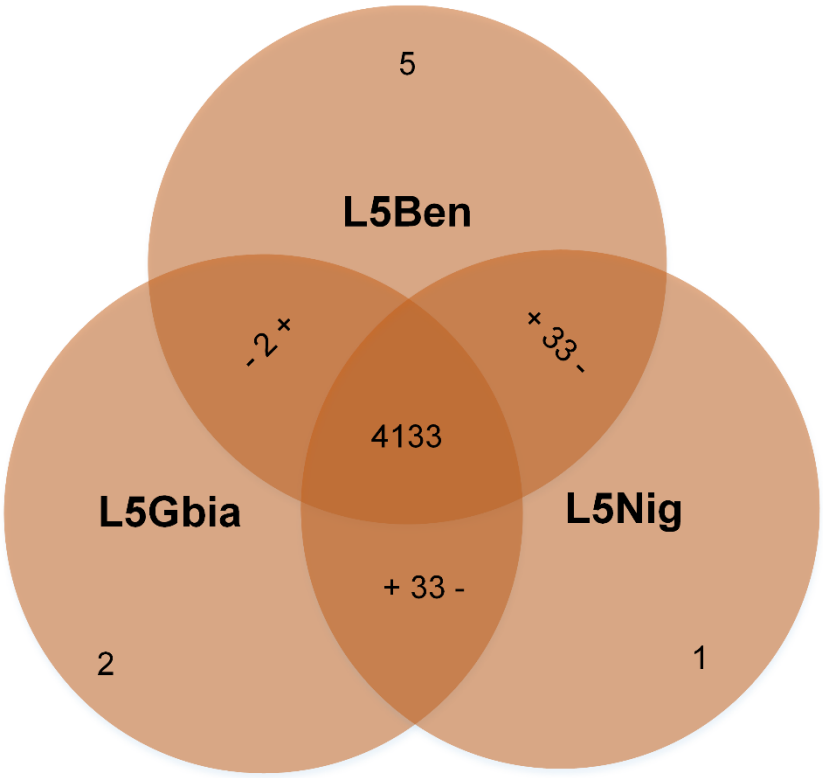


Figure 3. Gene count difference between three completed PacBio genomes of *M. africanum* L5 from Benin, The Gambia and Nigeria: L5Ben, L5Gbia and L5Nig. The numbers are the count/number of gene difference (present or absent) between genomes. The sign (+) indicates the presence of genes while (-) indicates the absence of genes. For example the "- 2 +" gene difference between L5Gbia and L5Ben, indicates that 2 genes are present in the L5Ben genome but absent in the L5Gbia genome. 5, 2, 1 are the number of genes only found in respectively L5Ben, L5Gbia, L5Nig.

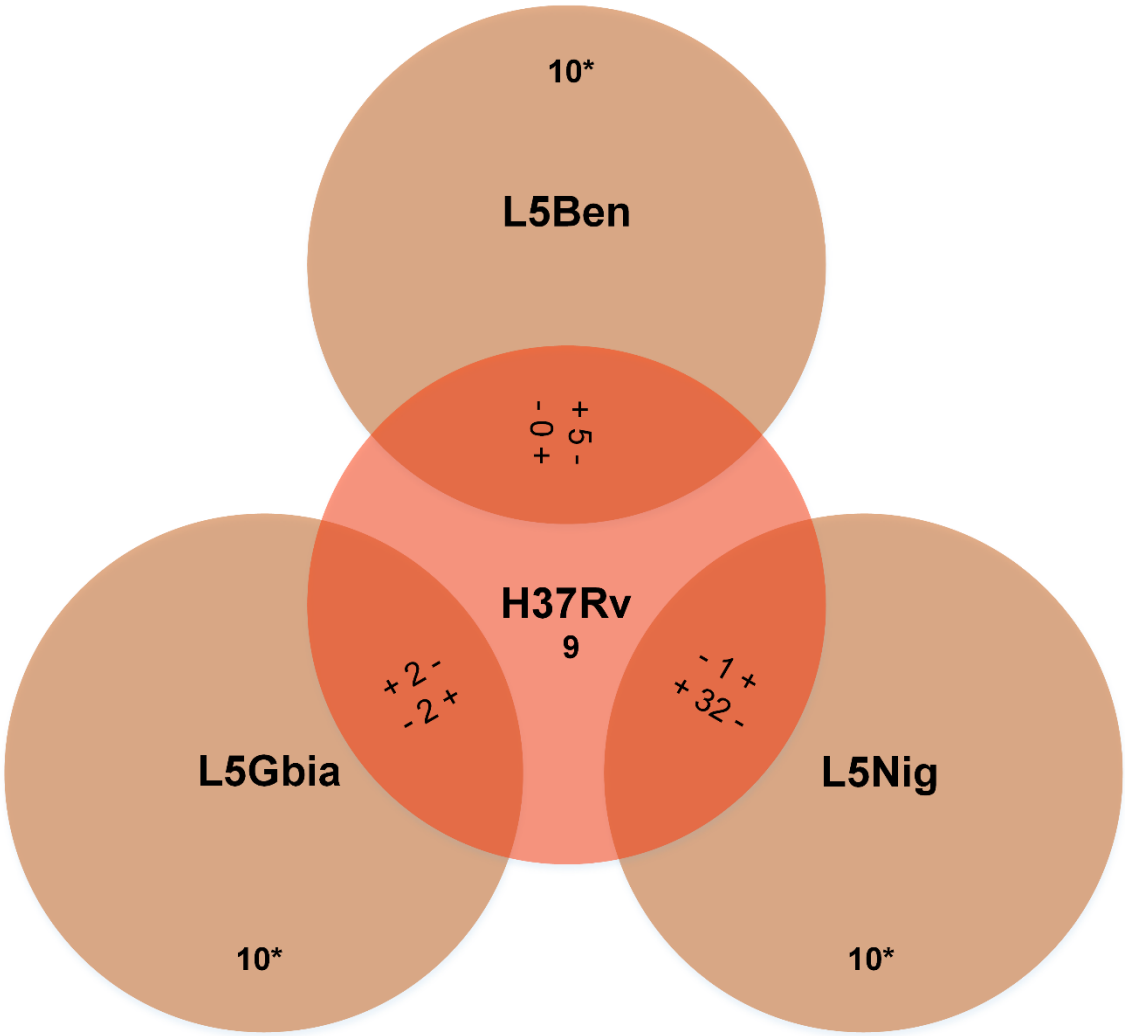
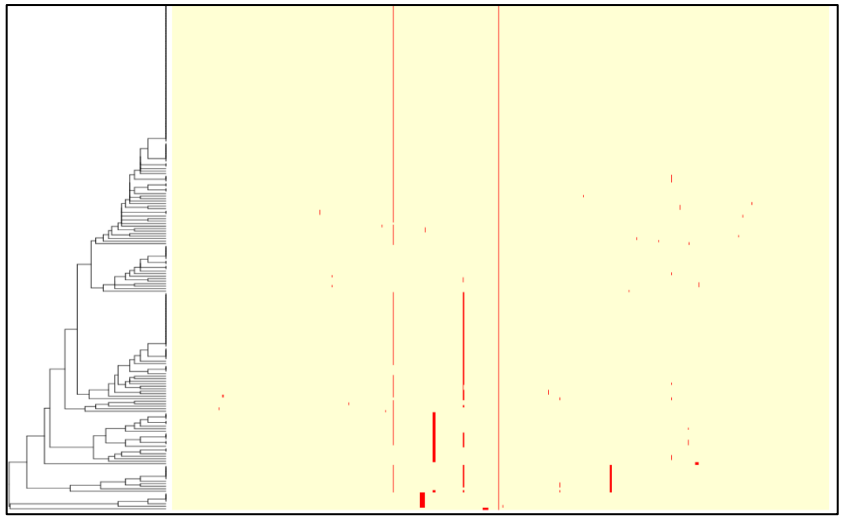
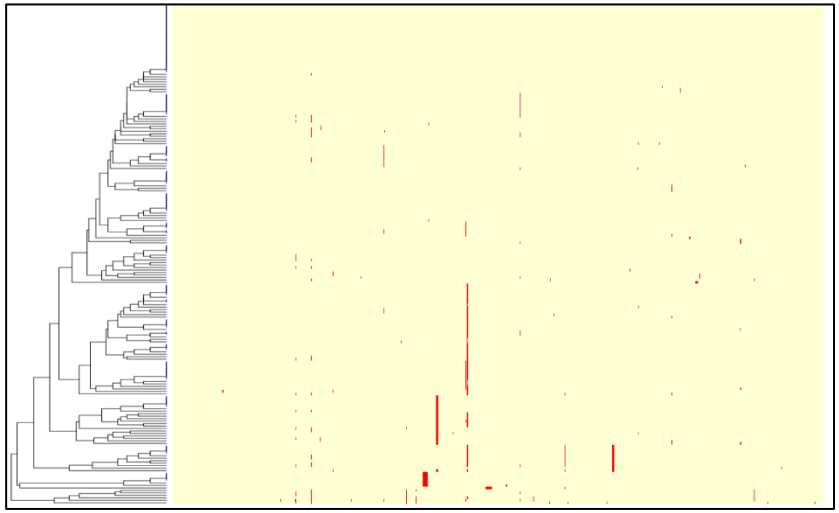


Figure 4. Gene count difference between *M. tuberculosis* H37Rv (L4) genome and three completed PacBio genomes of *M. africanum* L5 isolates from Benin, The Gambia and Nigeria. * These 10 genes are present in each of the three L5 completed PacBio genomes, and absent in H37RV.

H37Rv (L4) genes [A*]



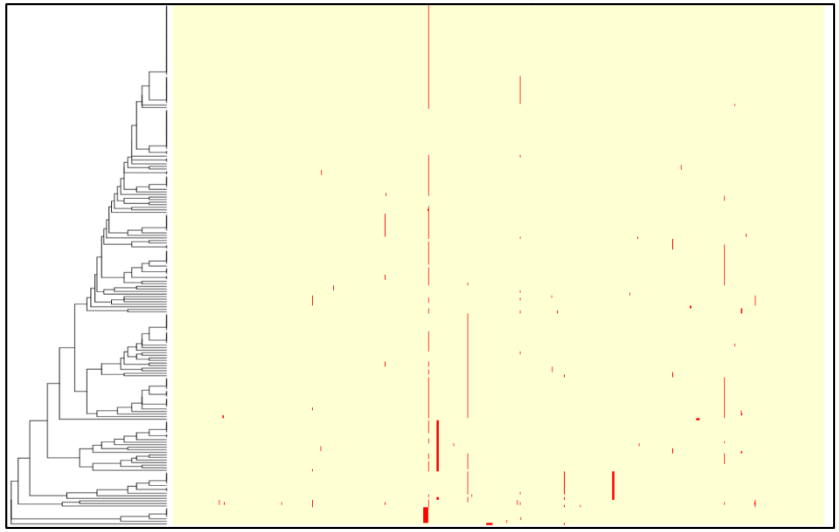
PcbL5Ben genes [B*]



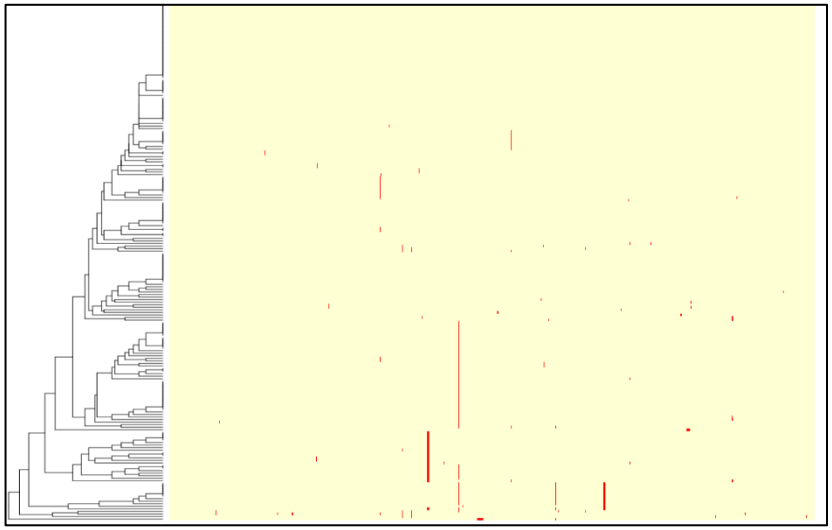
L5 Illumina genomes

942

PcbL5Gbia genes [C*]



PcbL5Nig genes [D]



L5 Illumina genomes

952 **Figure 5. Mapping of 205 clinical (Illumina platform sequenced) genomes (L5_ig) of *M. africanum* L5 from various countries to *M. tuberculosis* H37Rv (L4)**
953 **and three *M. africanum* L5 isolates (completed PacBio genomes) from Benin, The Gambia and Nigeria as reference (L5_cg_Ben; L5_cg_Gbia; L5_cg_Nig).**

954 (A) **5**/9 genes present in H37Rv but absent in the three L5_cg were **absent in all 205 L5_ig.**
955 (B) **5**/5 genes **present in L5_cg_Ben but absent in the two other L5_cg** were **present in 159-199/205 L5_ig.**
956 (C) **2**/2 genes **present in L5_cg_Gbia but absent in the two other L5_cg** were **present in 195 & 199/205 L5_ig.**
957 (D) **1**/1 gene **present in L5_cg_Nig but absent in the two other L5_cg** was **present in 195/205 L5_ig.**

958 * The 32-33 genes absent in L5_cg_Nig but present in L5_cg_Ben, L5_cg_Gbia and *H37Rv* were also absent in 6/205 L5_ig (Fig.9).

959 **.... or ....** gene absent

960

961

962

963

964

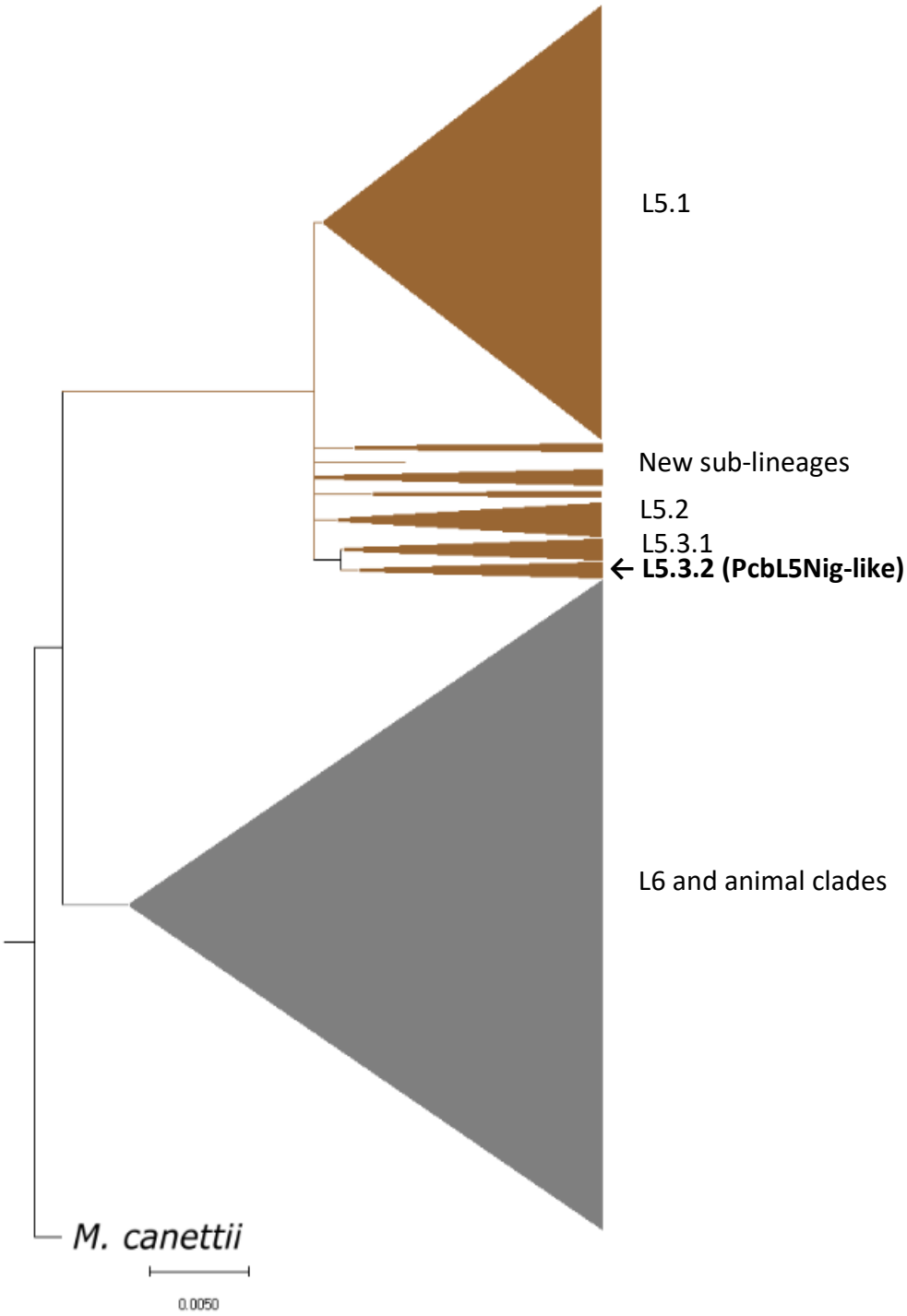
965

966

967

968

969



970

971

972 **Figure 6. Phylogenetic tree showing the 6 L5 isolates Illumina genomes (L5.3.2) similar to the**
973 **completed PacBio genome of the Nigerian L5 strain (PcbL5Nig).**

974

975 **Table 1. Gene content difference between *M. tuberculosis* H37Rv (L4) genome and completed**
976 **PacBio genomes of three L5 isolates from Benin, The Gambia and Nigeria.**

		Present in			
		PcbL5Ben	**PcbL5Gbia**	**PcbL5Nig**	**H37Rv**
Absent in	**PcbL5Ben**		2^{B}	1^{C}	**9^{D}**
	PcbL5Gbia	$2 + 5^{A}$		1^{C}	$2 + \mathbf{9^{D}}$
	PcbL5Nig	$33 + 5^{A}$	$33 + 2^{B}$		$32 + \mathbf{9^{D}}$
	H37Rv	$\mathbf{10^{*}} + 5^{A}$	$\mathbf{10^{*}} + 2^{B}$	$\mathbf{10^{*}} + 1^{C}$	
		4189	4162	4134	4126

977

978 **A:** include 5 only present in PcbL5Ben

979 **B:** includes 2 only present in PcbL5Gbia

980 **C:** includes 1 only present in PcbL5Nig

981 **D:** includes 9 only present in H37Rv

982 *: includes 10 genes shared by the PcbL5 (Ben, Gbia, Nig) and absent in H37Rv

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000 **Table 2. Mapping of 205 L5 isolates Illumina genomes to the *M. tuberculosis* H37Rv (L4) genome**
1001 **and completed genomes of three L5 isolates from Benin, The Gambia and Nigeria (mapping**
1002 **statistics/estimates)**

	H37Rv	**PcbL5Ben**	**PcbL5Gbia**	**PcbL5Nig**
Reads				
Mean of Percent L5 reads mapped to	96.9	**97.5**	97.3	96.5
Mean of unambiguous coverage mean	121.9	122.7	**122.8**	122.5
Bases				
Mean of Percent unambiguous total bases	0.983	0.988	**0.989**	0.982
Mean uncovered	30444.8	18801.7	**14657.5**	35215.4
Mean SNP	2210	529.9	**513.2**	**503.7**
Mean deletions	373.3	95.9	**95.7**	97.2
Mean insertions	239.2	**77.3**	107.6	**60.8**
Mean substitutions (including stop codons)	1193	**0**	**0**	**0**
Genes				
Mean of Percent gene mapped (presence)	99.59	99.72	**99.79**	99.76
L5 illumina having all the ref. genome genes % (n=205)	0	2 (4)	**7.3 (15)**	0
Mean gene count difference between L5 Illumina genomes (gene count per Illumina-sequenced genome minus minimum gene count)	29.8	38.3	**33.4**	33

1003

1004

1005 **Table 3. Presence in 205 L5 isolates Illumina genomes of genes detected in only one of the three L5**
1006 **completed genomes from Benin, Nigeria or The Gambia**

1007

		Present in L5 Illumina genomes % (n=205)
	PcbL5_01893	**77.6** (159)
	PcbL5_01894	**77.6** (159)
PcbL5Ben	PcbL5_01895	**77.6** (159)
	PcbL5_02043	**95.1** (195)
	PcbL5_03043	**97.1** (199)
PcbL5Gbia	PcbL5_02028	**95.1** (195)
	PcbL5_03020	**97.1** (199)
PcbL5Nig	PcbL5_02001	**95.1** (195)

1008

1009 **Table S1. Gene difference within the three L5 completed PacBio genomes from Benin, Nigeria or**
1010 **The Gambia, and between these three genomes and *M. tuberculosis* H37Rv**

Genes	Present in PcbL5Ben	Present in PcbL5Gbia	Present in PcbL5Nig	Present in H37Rv
Unshared (present in that genome only)	PcbL5Ben_01893, PcbL5Ben_01894, PcbL5Ben_01895, PcbL5Ben_02043, PcbL5Ben_03043	PcbL5Gbia_02028, PcbL5Gbia_03020	PcbL5Nig_02001	**H37Rv_02002, H37Rv_02085, H37Rv_02086, H37Rv_02087, H37Rv_02107, H37Rv_02108, H37Rv_02109, H37Rv_02188, H37Rv_02189**
Absent in PcbL5Ben		PcbL5Gbia_unshared	PcbL5Nig_unshared	H37Rv_unshared
Suspected pseudogenes in PcbL5Ben		PcbL5Gbia_00036, PcbL5Gbia_00955, PcbL5Gbia_02045, PcbL5Gbia_03308, PcbL5Gbia_03376	PcbL5Nig_00011, PcbL5Nig_00036, PcbL5Nig_00953, PcbL5Nig_01794, PcbL5Nig_01853, PcbL5Nig_01896, PcbL5Nig_02018, PcbL5Nig_02119, PcbL5Nig_02291, PcbL5Nig_02467	H37Rv_00012, H37Rv_00036, H37Rv_00528, H37Rv_01371, H37Rv_01801, H37Rv_02020, H37Rv_02463, H37Rv_02556, H37Rv_02951, H37Rv_03349, H37Rv_03797
Absent in PcbL5Gbia	PcbL5Ben_01443, PcbL5Ben_01444, PcbL5Ben_unshared	PcbL5Gbia_unshared	PcbL5Nig_unshared	H37Rv_01413, H37Rv_01414, H37Rv_unshared
Suspected pseudogenes in PcbL5Gbia	PcbL5Ben_01064, PcbL5Ben_02378, PcbL5Ben_02486, PcbL5Ben_03113		PcbL5Nig_00011, PcbL5Nig_01794, PcbL5Nig_01853, PcbL5Nig_01896, PcbL5Nig_02119, PcbL5Nig_02291, PcbL5Nig_02467	H37Rv_00012, H37Rv_00528, H37Rv_01371, H37Rv_01801, H37Rv_02329, H37Rv_02463, H37Rv_02556, H37Rv_02951, H37Rv_03797,
Absent* in PcbL5Nig	PcbL5Ben_01617, PcbL5Ben_01618, PcbL5Ben_01619, PcbL5Ben_01620, PcbL5Ben_01621, PcbL5Ben_01622, PcbL5Ben_01623, PcbL5Ben_01624, PcbL5Ben_01625, PcbL5Ben_01626, PcbL5Ben_01627, PcbL5Ben_01628,	PcbL5Gbia_01604, PcbL5Gbia_01605, PcbL5Gbia_01606, PcbL5Gbia_01607, PcbL5Gbia_01608, pcbL5Gbia_01609, PcbL5Gbia_01610, PcbL5Gbia_01611, PcbL5Gbia_01612, pcbL5Gbia_01613, PcbL5Gbia_01614, pcbL5Gbia_01615,		H37Rv_01581, H37Rv_01582, H37Rv_01583, H37Rv_01584, H37Rv_01585, H37Rv_01586, H37Rv_01587, H37Rv_01588, H37Rv_01589, H37Rv_01590, H37Rv_01591, H37Rv_01592,

	PcbL5Ben_01629, PcbL5Ben_01630, PcbL5Ben_01631, PcbL5Ben_01632, PcbL5Ben_01633, PcbL5Ben_01634,	PcbL5Gbia_01616, PcbL5Gbia_01617, PcbL5Gbia_01618, pcbL5Gbia_01619, PcbL5Gbia_01620, PcbL5Gbia_01621,		H37Rv_01593, H37Rv_01594, H37Rv_01595, H37Rv_01596, H37Rv_01597, H37Rv_01598,
	PcbL5Ben_01636, PcbL5Ben_01637, PcbL5Ben_01638, PcbL5Ben_01639, PcbL5Ben_01640, PcbL5Ben_01641, PcbL5Ben_01642, PcbL5Ben_01643, PcbL5Ben_01644, PcbL5Ben_01645, PcbL5Ben_01646, PcbL5Ben_01647,	PcbL5Gbia_01623, PcbL5Gbia_01624, PcbL5Gbia_01625, pcbL5Gbia_01626, PcbL5Gbia_01627, PcbL5Gbia_01628, PcbL5Gbia_01629, PcbL5Gbia_01630, PcbL5Gbia_01631, PcbL5Gbia_01632, pcbL5Gbia_01633, PcbL5Gbia_01634,		H37Rv_01600, H37Rv_01601, H37Rv_01602, H37Rv_01603, H37Rv_01604, H37Rv_01605, H37Rv_01606, H37Rv_01607, H37Rv_01608, H37Rv_01609, H37Rv_01610,
	PcbL5Ben_01891, PcbL5Ben_01892, PcbL5Ben_02248, PcbL5Ben_unshared	PcbL5Gbia_01881, PcbL5Gbia_01882, PcbL5Gbia_02228, PcbL5Gbia_unshared		H37Rv_01853, H37Rv_01854, H37Rv_02202 H37Rv_unshared
Suspected pseudogenes in PcbL5Nig	PcbL5Ben_01064, PcbL5Ben_01718, PcbL5Ben_02486	PcbL5Gbia_01705, PcbL5Gbia_03308, PcbL5Gbiaa_03376		H37Rv_00528, H37Rv_01371, H37Rv_01679, H37Rv_02463, H37Rv_02556, H37Rv_02951, H37Rv_03349, H37Rv_03797
Absent in H37Rv	**PcbL5Ben_01364,** **PcbL5Ben_01365,** **PcbL5Ben_02129,** **PcbL5Ben_02130,** **PcbL5Ben_02181,** **PcbL5Ben_02182,** **PcbL5Ben_03555,** **PcbL5Ben_03556,** **PcbL5Ben_03557,** **PcbL5Ben_03669** PcbL5Ben_unshared	**PcbL5Gbia_01359,** **PcbL5Gbia_01360,** **PcbL5Gbiaa_02112,** **PcbL5Gbia_002113,** **PcbL5Gbia_02163,** **PcbL5Gbiaa_02164,** **PcbL5Gbia_03527,** **PcbL5Gbia_03528,** **PcbL5Gbiaa_03529,** **PcbL5Gbia_03643,** PcbL5Gbia_unshared	**PcbL5Nig_01357,** **PcbL5Nig_01358,** **PcbL5Nig_02085,** **PcbL5Nig_02086,** **PcbL5Nig_02137,** **PcbL5Nig_02138,** **PcbL5Nig_03500,** **PcbL5Nig_03503,** **PcbL5Nig_03504,** **PcbL5Nig_03616,** PcbL5Nig_unshared	
Suspected pseudogenes in H37Rv	PcbL5Ben_00749, PcbL5Ben_00828, PcbL5Ben_01064, PcbL5Ben_01108, PcbL5Ben_02128, PcbL5Ben_02486, PcbL5Ben_03113,	PcbL5Gbia_00750, PcbL5Gbia_00826, PcbL5Gbiaa_00955, PcbL5Gbia_01103, PcbL5Gbia_01879, PcbL5Gbiaa_02111, PcbL5Gbia_03117,	PcbL5Nig_00748, PcbL5Nig_00824, PcbL5Nig_00953, PcbL5Nig_01100, PcbL5Nig_01846, PcbL5Nig_01853, PcbL5Nig_01896, PcbL5Nig_02084,	

	PcbL5Ben_03141, PcbL5Ben_03667	PcbL5Gbia_03308, PcbL5Gbiaa_03641	PcbL5Nig_02119, PcbL5Nig_02291, PcbL5Nig_02467, PcbL5Nig_03064, PcbL5Nig_03092, PcbL5Nig_03614

*Genes hilighted in the same color are contiguous and form a block.

1038 **Table S2. Verification of genes in the L5 completed PacBio genomes suspected to be pseudogenes**
1039 **in *M. tuberculosis* H37Rv genome.**

Gene	Gene in PcbL5Ben	Gene in PcbL5Gbia	Gene in PcbL5Nig	Length (aa)	Locus in H37Rv	Remarks	Conclusion
1	**PcbL5Ben_00749**	**PcbL5Gbia_00750**	**PcbL5Nig_00748**	101	798857-799150	M1V, L96p, H97t, del NCD (aa 99-101)	**Pseudogene in H37Rv**
2	PcbL5Ben_00828	PcbL5Gbia_00826	PcbL5Nig_00824	34	863155-863256	No change	Not pseudogene in H37Rv
3	Not suspected pseudogene	PcbL5Gbia_00955	PcbL5Nig_00953	204	993212-992602	Inverted, M1V	**Pseudogene in H37Rv**
4	PcbL5Ben_01064	Not suspected pseudogene	Not suspected pseudogene	34	1102546-1102647	S23r	Not pseudogene in H37Rv
5	**PcbL5Ben_01108**	**PcbL5Gbia_01103**	**PcbL5Nig_01100**	70	1148218-1148397	M1V	**Pseudogene in H37Rv**
6	**PcbL5Ben_01889**	**PcbL5Gbia_01879**	**PcbL5Nig_01846**	68	1981341-1981341	M1V	**Pseudogene in H37Rv**
7	**PcbL5Ben_02128**	**PcbL5Gbia_02111**	**PcbL5Nig_02084**	92	2219418-2219318	Inverted, alignment showed aa 59-92, no mutation	**Pseudogene in H37Rv**
8	Not suspected pseudogene	Not suspected pseudogene	PcbL5Nig_01853 is PcbNig_1896	41	2038813-2038706	Inverted, alignment showed aa 59-92, no mutation	**Pseudogene in H37Rv**
9	PcbL5Ben_02486	Not suspected pseudogene	Not suspected pseudogene	41	2582440-2582318	Inverted, M1V	**Pseudogene in H37Rv**
10	**PcbL5Ben_03141**	**PcbL5Gbia_03117**	**PcbL5Nig_03092**	111	3291532-3291350	Inverted, del aa 1-aa60 (60 first aa)	**Pseudogene in H37Rv**
11	Not suspected pseudogene	Not suspected pseudogene	PcbL5Nig_02119	46	2256620-2256483	Inverted, No change	**Pseudogene in H37Rv**
12	Not suspected pseudogene	Not suspected pseudogene	PcbL5Nig_02291	33	24345567-2434665	No change	Not pseudogene in H37Rv
13	Not suspected pseudogene	Not suspected pseudogene	PcbL5Nig_02467	94	2614352-2614071	Inverted	**Pseudogene in H37Rv**
14	PcbL5Ben_03113	Not suspected pseudogene	PcbL5Nig_03060	204	3232652-3232867	Del aa1-aa56, N74D, del aa129-aa204	**Pseudogene in H37Rv**
15	Not suspected pseudogene	PcbL5Gbia_03308	Not suspected pseudogene	58	3487871-3487698	Inverted, M1V	**Pseudogene in H37Rv**
16	**PcbL5Ben_03667**	**PcbL5Gbia_03641**	**PcbL5Nig_03614**	99	3841840-3842109	M1V, all aa mutated (insertion substitution) from aa26-aa81, del aa82-aa98	**Pseudogene in H37Rv**

1040

1041 **Table S3. Verification of genes present in H37Rv suspected to be pseudogenes in the L5 completed**
1042 **genomes**

Gene in H37Rv	Length (aa)	Locus in H37Rv	Locus in PcbL5Ben	Locus in PcbL5Gbia	Locus in PcbL5Nig	Remarks	Conclusion
H37Rv_02463	42	2604950-2605078	2631987-2632112	2614961-2615086	2603370-2603495	No change	Not pseudogene
H37Rv_02556	67	2718844-2719047	2743020-2742820	2729937-2729737	2716988-2716788	Inverted, nSNP I66L	**Pseudogene in the 3 Pacbio L5**
H37Rv_02951	55	3115741-3115908	3137926-3137762	3126133-3125969	3112866-3112702	Inverted, nSNP M1V, L55F	**Pseudogene in the 3 Pacbio L5**
H37Rv_03797	200	4053036-4053638	4076665-4077105	4064716-4065156	4055914-4056354	del aa 1-53	**Pseudogene in the 3 Pacbio L5**

1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062

1063 **Table S4. Genes absent in the Nigerian completed PacBio genome (PcbL5Nig) and some L5 isolates Illumina genomes (n=6), but present in the Benin and**
1064 **Gambian completed PacBio genomes (PcbL5Ben, PcbL5Gbia), H37Rv and all other L5 isolates Illumina genomes.**

ID (this study)	Rv number	Functional group (Mycobrowser)	Gene name (Mycobrowser)	Product (Mycobrowser)	Function (Mycobrowser)
H37Rv_01581	*Rv1493*	Lipid metabolism	*mutB*	Probable methylmalonyl-CoA mutase large subunit MutB (MCM)	Involved in propionic acid fermentation
H37Rv_01582	*Rv1494*	Virulence, detoxification, adaptation	*mazE4*	Possible antitoxin MazE4	Possible mazE4, antitoxin, part of toxin-antitoxin (TA) operon with Rv1495. Non-essential gene for in vitro growth of H37Rv
H37Rv_01583	*Rv1495*	Virulence, detoxification, adaptation	*mazF4*	Possible toxin MazF4	Sequence-specific mRNA cleavage
H37Rv_01584	*Rv1496*	Cell wall and cell processes	*Rv1496*	Possible transport system kinase	Possibly involved in transport (possibly arginine)
H37Rv_01585	*Rv1497*	Intermediary metabolism and respiration	*lipL*	Probable esterase LipL	Function unknown, but supposed involvement in lipid metabolism
H37Rv_01586	*Rv1498A*	Conserved hypotheticals	*Rv1498A*	Conserved protein	Function unknown
H37Rv_01587	*Rv1498c*	Intermediary metabolism and respiration	*Rv1498c*	Probable methyltransferase	Causes methylation
H37Rv_01588	*Rv1499*	Conserved hypotheticals	*Rv1499*	Hypothetical protein	Function unknown
H37Rv_01589	*Rv1500*	Intermediary metabolism and respiration	*Rv1500*	Probable glycosyltransferase	Function unknown
H37Rv_01590	*Rv1501*	Conserved hypotheticals	*Rv1501*	Conserved hypothetical protein	Function unknown
H37Rv_01591	*Rv1502*	Unknown	*Rv1502*	Hypothetical protein	Function unknown
H37Rv_01592	*Rv1505c*	Conserved hypotheticals	*Rv1505c*	Conserved hypothetical protein	Function unknown. It has some similarity to hypothetical proteins and glycosylases
H37Rv_01593	*Rv1506c*	Unknown	*Rv1506c*	Hypothetical protein	Function unknown. Non-essential gene for in vitro growth of H37Rv

H37Rv_01594	*Rv1507A*				
H37Rv_01595	*Rv1507c*	Conserved hypotheticals	*Rv1507A*	Hypothetical protein	Function unknown
H37Rv_01596	*Rv1508A*	Conserved hypotheticals	*Rv1508A*	Conserved hypotheticals	Function unknown. Highly similar to central part of glycosyl transferases from various mycobacteria and eubacteria
H37Rv_01597	*Rv1508c*	Cell wall and cell processes	*Rv1508c*	Probable membrane protein	Function unknown. Predicted to be in the GT-C superfamily of glycosyltransferases
H37Rv_01598	*Rv1509*	Unknown	*Rv1509*	Hypothetical protein	Function unknown
H37Rv_01599	*Rv1510* (present in PcbL5Nig but missing in the PcbL5Nig-like Illumina L5)	Cell wall and cell processes	*Rv1510*	Probable conserved membrane protein	Function unknown
H37Rv_01600	*Rv1511*	Intermediary metabolism and respiration	*gmdA*	GDP-D-mannose dehydratase GmdA (GDP-mannose 4,6 dehydratase)	Function unknown, probably involved in nucleotide-sugar metabolism
H37Rv_01601	*Rv1512*	Intermediary metabolism and respiration	*epiA*	Probable nucleotide-sugar epimerase EpiA	Function unknown, probably involved in nucleotide-sugar metabolism
H37Rv_01602	*Rv1513*	Conserved hypotheticals	*Rv1513*	Conserved protein	Function unknown, similar to hypothetical proteins from several organisms
H37Rv_01603	*Rv1514c*	Conserved hypotheticals	*Rv1514c*	Conserved hypothetical protein	Function unknown; similar to other hypothetical protein and to putative colanic acid biosynthesis glycosyl transferase
H37Rv_01604	*Rv1515c*	Conserved hypotheticals	*Rv1515c*	Conserved hypothetical protein	Function unknown
H37Rv_01605	*Rv1516c*	Intermediary metabolism and respiration	*Rv1516c*	Probable sugar transferase	Function unknown, involved in cellular metabolism. Non essential for in vitro growth.
H37Rv_01606	*Rv1517*	Cell wall and cell processes	*Rv1517*	Conserved hypothetical membrane protein	Function unknown

H37Rv_01607	*Rv1518*	Conserved hypotheticals	*Rv1518*	Conserved hypothetical protein	possibly glycosyl transferase involved in exopolysaccharide synthesis, similar to several hypothetical proteins and glycosyl transferases from diverse organisms
H37Rv_01608	*Rv1519*	Conserved hypotheticals	*Rv1519*	Conserved hypothetical protein	Function unknown
H37Rv_01609	*Rv1520*	Intermediary metabolism and respiration	*Rv1520*	Probable sugar transferase	Funtion unknown; thought to be involved in cellular metabolism
H37Rv_01610	*Rv1521*	Lipid metabolism	*fadD25*	Probable fatty-acid-AMP ligase FadD25 (fatty-acid-AMP synthetase) (fatty-acid-AMP synthase)	Function unknown; involved in lipid degradation
	Rv1522c (present in PcbL5Nig but missing in the PcbL5Nig-like Illumina L5)	Cell wall and cell processes	*mmpL12*	Probable conserved transmembrane transport protein MmpL12	Function unknown. Thought to be involved in fatty acid transport.

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076 **Table S5. Presence /absence in 205 L5 isolates Illumina genomes of the 10 genes shared by the three L5 completed genomes but absent in the *M.***
1077 ***tuberculosis* H37Rv genome.**

Genes present in L5 completed PacBio genomes but absent in H37Rv	Present in L5 isolates illumina genome n=205 (%)	SNP in L5-specific genes			Gene function			Present in complete genome of
		PcbL5Ben as ref n=205 (%)	PcbL5Gbia as ref n=205 (%)	PcbL5Nig as ref n=205 (%)	Functional group (Mycobrowser)	genes	Summary of role based on the literature	
PcbL5Ben_1364	191 (93.2)				ABC transporter ATP-binding protein/permease (Prokka)	Transmembrane ATP-binding protein ABC transporter	Conserved domains: FHA and CcmA (NCBI) FHA (forkhead associated domain, COG1716 (T)) binds pSer, pThr, pTyr (Signal transduction mechanisms) CcmA (COG1131 (V)), ABC-type multidrug transport system, ATPase component (Defense mechanisms (NCBI))	L5, bovis
PcbL5Ben_1365	191 (93.2)				Same as PcbL5Ben_1364	Same as PcbL5Ben_1364	Same as PcbL5Ben_1364	L5, bovis
PcbL5Ben_2129#	**205 (100)**	4 (2)	4 (2)	205* (100)	**PE/PPE**	PPE, possibly PE35 (77% id with PE35 from *M. braziliense*)	PE35 implicated in virulence (STRING database) and PE family is suggested to be related to antigenic variation of MTBC (NCBI)	**Only L5**
PcbL5Ben_2130#	**205 (100)**	7 (3.4)	7 (3.4)	7 (3.4)	Not in Mycobrowser	**Hypothetical protein possibly CAAX conserved domain** (NCBI)	Involved in post-translation modification by attaching to the isoprenoid proteins in the process called prenylation (NCBI). Most of CAAX box proteins do not have a transmembrane domain, thus, the prenylation process is crucial for the function of many signal transduction proteins (Gao, Liao, and Yang 2009a)	**Only L5**
PcbL5Ben_2181	**205 (100)**	3 (1.5)	3 (1.5)	3 (1.5)	Not in Mycobrowser	**Hypothetical protein,**	IS256 transposase is implicated in resistance to antibiotics and virulence (recombination	L5, L6, bovis

						possibly IS256 transposase (90% id, NCBI)	for adaptation, invasion)(Murugesan et al. 2018; Gu et al. 2005)	
PcbL5Ben_2182	**205 (100)**	6 (2.9)	6 (2.9)	6 (2.9)	**Unknown function (Mb2048c**)	*Mb2048c*	Function unknown (expressed during exponential growth in Sauton's minimal media (NCBI)). Gene absent in H37Rv (Mycobrowser)	L5, L6, bovis
PcbL5Ben_3555	167 (81.5)				Intermediary metabolism and respiration	moaA3 (NCBI)	*moaA3* (molybdenum cofactor biosynthesis protein subunit MoaA , Cyclic pyranopterin monophosphate synthase (Prokka)) is contained in an IS6110 sequence deleted in H37Rv (Mycobrowser)	L5, L6, bovis
PcbL5Ben_3556	165 (80.5)				Not in Mycobrowser	Hypothetical protein, possibly DnrI superfamily (NCBI)	*DnrI*: DNA binding transcriptional activator of the SARP family (signal transduction mechanisms, COG3629 (NCBI))	L5, L6, bovis
PcbL5Ben_3557	165 (80.5)				Peptide synthase (Prokka)	Hypothetical protein, possibly IS6110 transposase (99-100% id NCBI)		L5, L6, bovis
PcbL5Ben_3669	201 (98)				PE/PPE	Hypothetical protein		L5, L6

*In total 210 in the 205 L5 Illumina genomes: more than one SNP in the gene for some strains

Present in L5 only, not present in L6 nor *M. bovis.* Those two genes (PcbL5Ben_2129 and 2130) are adjacent.

PcbL5Ben_1364 and PcbL5Ben_1365 are adjacent genes

PcbL5Ben_2181 and PcbL5Ben_2182 are adjacent genes

PcbL5Ben_3555, PcbL5Ben_3556 and PcbL5Ben_3557 are adjacent genes

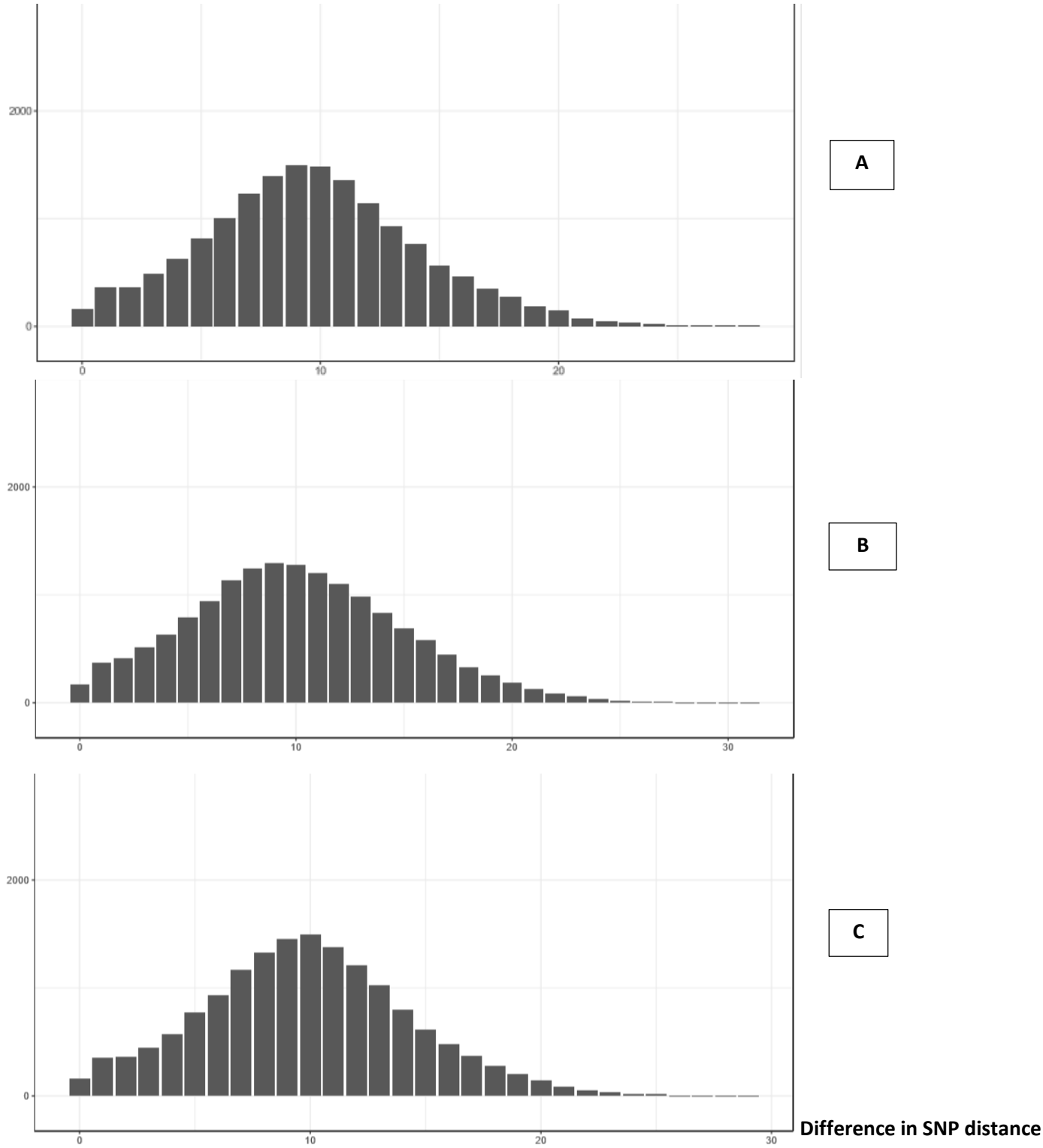
1087 **Table S6. Presence/absence in 205 Illumina L5 genomes of the nine genes present in the H37Rv genome but absent in the three completed L5 genomes.**

Genes present in H37Rv but absent in L5 completed PacBio genomes	Present in L5 isolates Illumina genomes n=205 (%)	Absence in L5 isolates illumina genomes n=205 (%)	Gene function			Absent in complete genome of	Present in complete genome of
			Functional group (Mycobrowser)	Genes	Summary of role based on the literature		
Part of Rv1899c (H37Rv_2002)	204 (99.5)	1 (0.5)	Hypothetical protein (35 aa) part of *Rv1899c* (343 aa, Cell wall and cell processes)(Tuberculist)	Hypothetical protein, part of Rv1899c (LppD: 4-hydroxybuturate dehydrogenase 173 aa) and also LpqI beta-hexoaminidase precursor (Herrmann et al. 2000)	Possibly bacterial membrane lipoprotein (LppD and LpqI)	L5, L6, bovis	**H37Rv only**
Rv1977 (H37Rv_2085)	**0**	**205 (100)**	**Conserved hypotheticals**	M48 peptidase family protein including as homologs: CAAX prenyl protease (*Htpx* (COG0501 O, heat shock stress response protein) M48-Ste24p-like	Bacterial survival (bacterial self-degradation to eliminate abnormal membrane proteins, htpx upregulated at high temperature, and also after 96 hours of starvation (NCBI))	**L5 only**	L6, bovis, H37Rv
Rv1978 (H37Rv_2086)	3 (1.5)	202 (98.5)	Conserved hypotheticals,	M48-2C-Ste24p (100% id), class I SAM-dependent methyltransferase (99.65% id), type 11 methyltransferase (97.1% id)	Bacterial survival in macrophages, and non-essential for in vitro growth of H37Rv (Rengarajan, Bloom, and Rubin 2005)(Mycobrowser, Tuberculist)	L5 only	L6, bovis, H37Rv
Rv1979c (H37Rv_2087)	**0**	**205 (100)**	**Cell wall and cell processes**	APC family permease	-Involved in the transport of **clofazimine** (and **bedaquiline**), associated with clofazimine and bedaquiline resistance (Zhang et al. 2015; Ghodousi et al. 2019)-Disruption of the gene provides an in vitro growth advantage to H37Rv	**L5 only**	L6, bovis, H37Rv

Rv1993c (H37Rv_2107)	**0**	**205 (100)**	**Conserved hypotheticals**	*Rv0968 =DUF1490* similar to *Rv1993c* (NCBI)	Similar to *Rv0968* which is included in the operon *cosR-Rv0968-ctpV* where *cosR* and *ctpV* are virulence associated (Rademacher and Masepohl 2012)	**L5 only**	L6, bovis, H37Rv
Rv1994c (H37Rv_2108)	1 (0.5)	204 (99.5)	Regulatory proteins	HTH transcriptional regulator *cmtR*	Operon *cmtR-Rv1993c-ctpG* is similar to *csoR-Rv0968-ctpV*. *cmtR* (*Rv1994c*) has the same role as *csoR*. These operons all function in the regulation and transport (efflux) of toxic metal especially copper which is toxic in excess (Samanovic et al. 2012). Disruption may hamper in vitro growth, and survival during chronic phase of infection as *csoR* absence (Marcus et al. 2016; Rowland and Niederweis 2012; Ward, Hoye, and Talaat 2008).	L5 only	L6, bovis, H37Rv
Rv1995 (H37Rv_2109)	**0**	**205 (100)**	**Conserved hypothetical**	**Hemerythrin-domain containing protein (NCBI)**	- Involved in oxygen transport (NCBI)	**L5 only**	L6, bovis, H37Rv
Rv2073c (H37Rv_2188)	**0**	**205 (100)**	**Intermediary metabolism and respiration**	**Probable short chain dehydrogenase**	SDR family NAD(P)-dependent oxidoreductase (NCBI): catalyzes a wide range of reactions and substrate(Kavanagh et al. 2008; Sellés Vidal et al. 2018)	L5, L6, bovis	**H37Rv only**
Rv2074 (H37Rv_2189)	1 (0.5)	204 (99.5)	**Intermediary metabolism and respiration (Pyridoxamine, Vit B6)**	Pyridoxamine-5-phosphate oxidase (Mycobrowser, Tuberculist, NCBI) Its cofactor (Selengut and Haft 2010) F420 dependent biliverdin reductase (99.2-100% id, NCBI)(Ahmed et al. 2016)	- Biosynthesis of (pyridoxal phosphate and) pyridoxine (vitamin B6, Mycobrowser) which is essential for survival and virulence of M. tuberculosis (Dick et al. 2010) - Its cofactor implicated in Immuno-evasive mechanism to allow bacterial persistence (Ahmed et al. 2016; Selengut and Haft 2010)	L5, L6, bovis	H37Rv only

1088

1089



1090

1091

1092

1093 **Figure S1. Comparison of reference genome (differences) for SNP distance between pairwise**
1094 **Illumina L5 genomes**. A: Difference between H37Rv and PacBio L5 Benin genome; B: Difference
1095 between H37Rv and PacBio L5 Gambian genome; C: Difference between H37Rv and PacBio L5
1096 Nigerian genome

1097

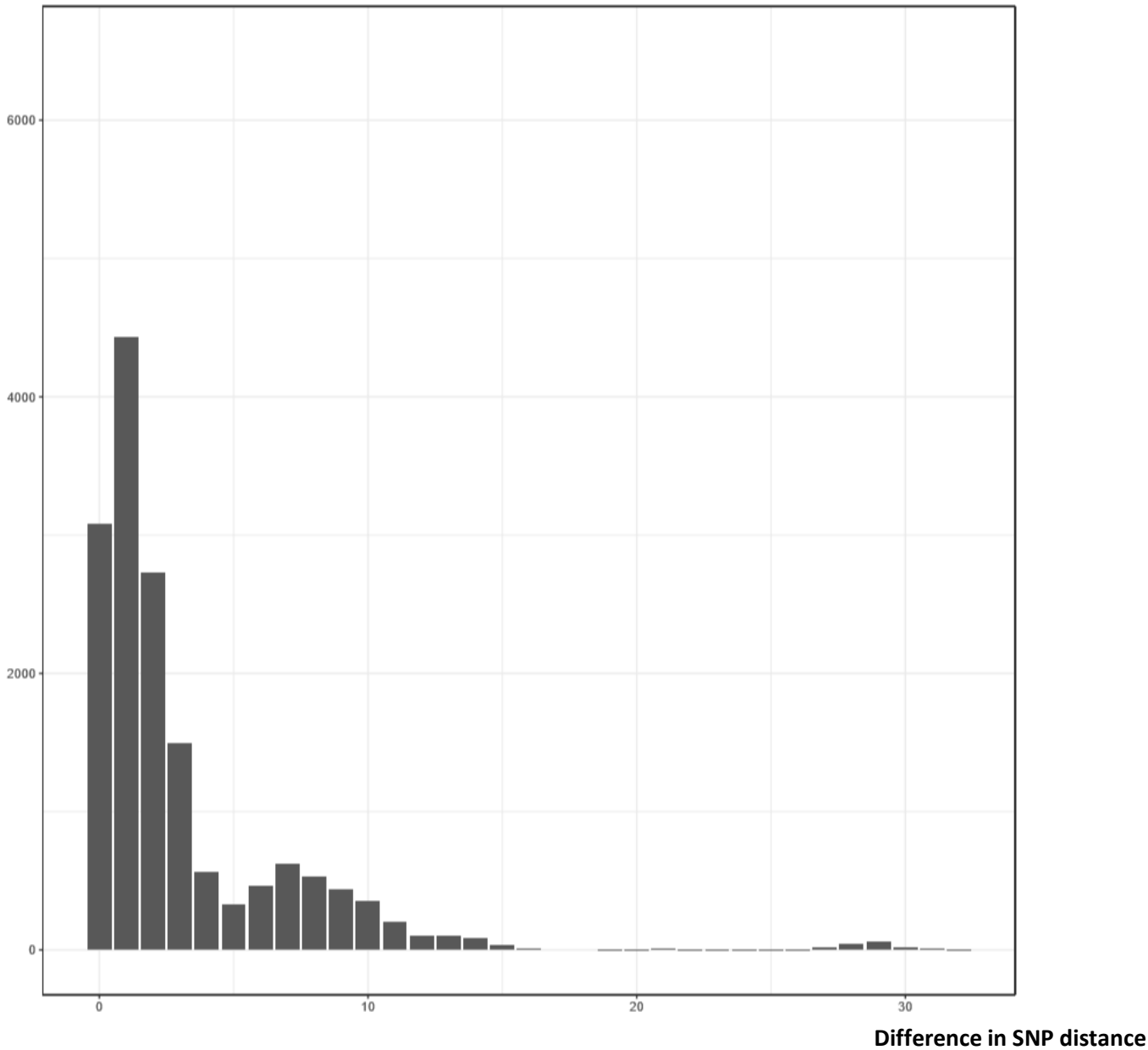
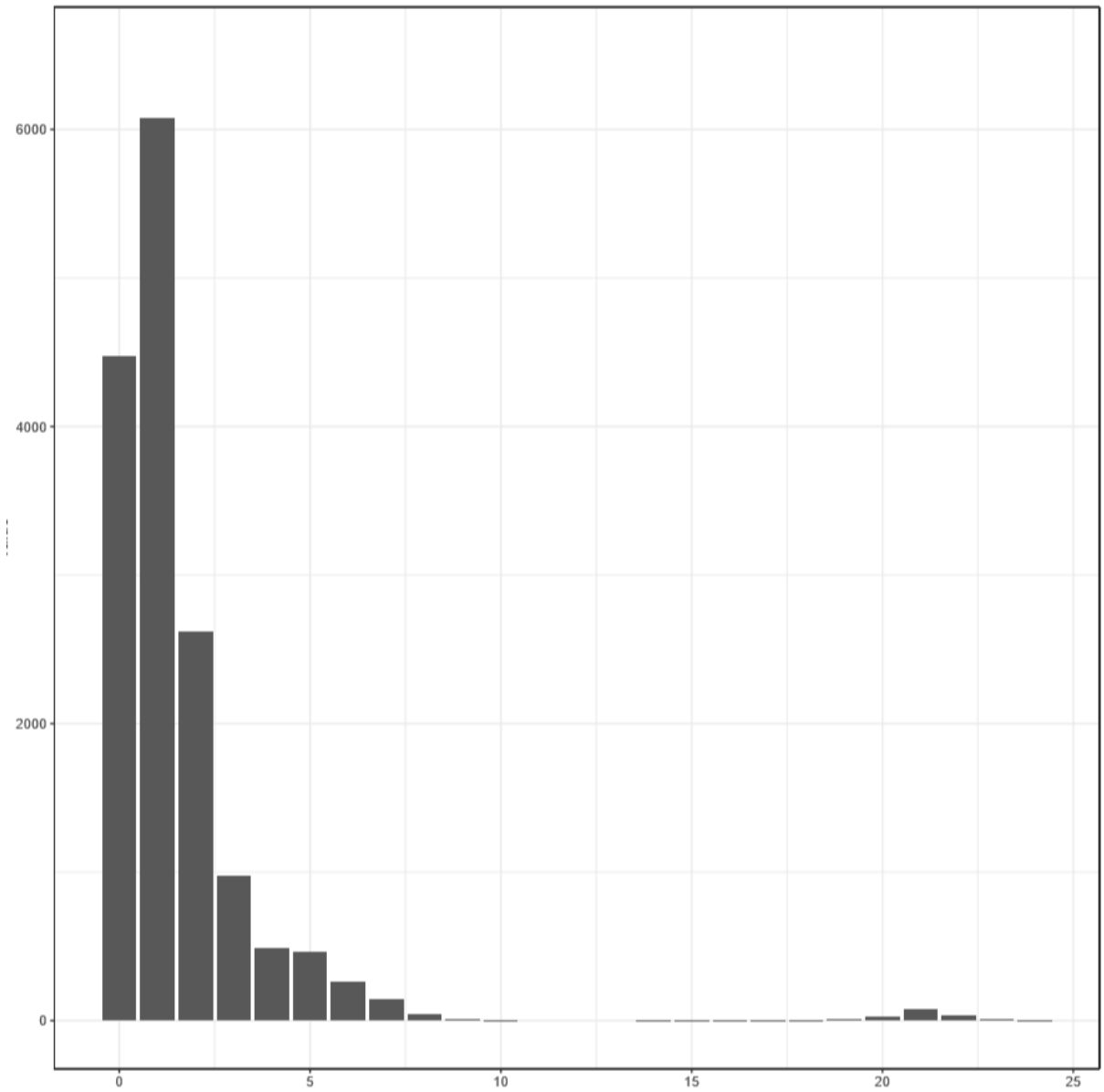


Figure S2-A. Comparison of reference genome (differences) for SNP distance between pairwise Illumina L5 genomes: difference between PacBio L5 Benin genome and PacBio L5 Gambian genome

1110

1111



1112

Difference in SNP distance

1113

1114 **Figure S2-B. Comparison of reference genome (differences) for SNP distance between pairwise**
1115 **Illumina L5 genomes:** difference between PacBio L5 Benin genome and PacBio L5 Nigerian genome

1116

1117

1118

1119

1120

1121

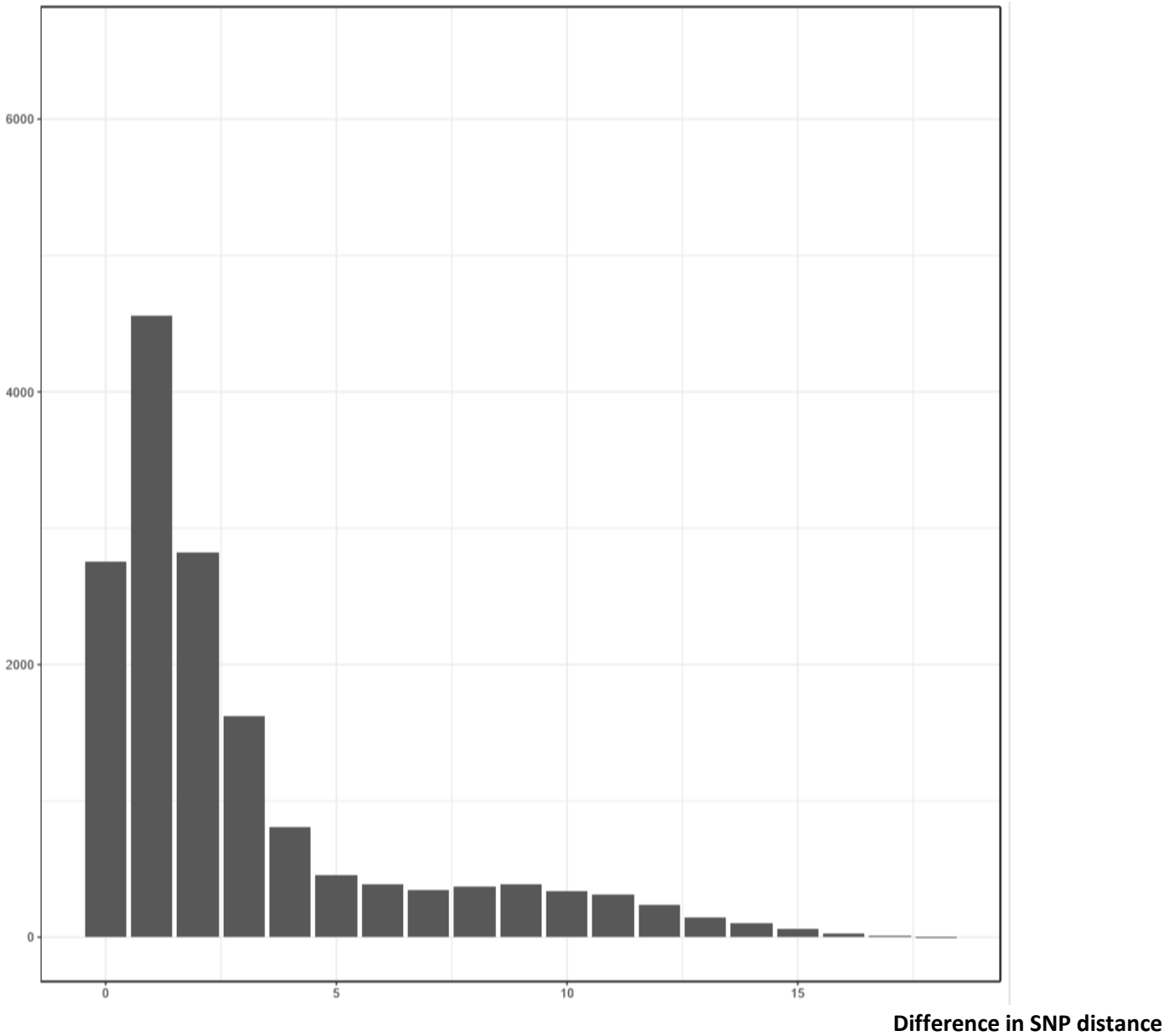


Figure S2-C. Comparison of reference genome (differences) for SNP distance between pairwise Illumina L5 genomes: difference between PacBio L5 Gambian genome and PacBio L5 Nigerian genome