

Towards next generation diagnostics for tuberculosis: identification of novel molecular targets by large-scale comparative genomics.

Galo A. Goig¹, Manuela Torres-Puente¹, Carla Mariner-Llicer¹, Luis M. Villamayor², Álvaro Chiner-Oms³, Ana Gil-Brusola⁴, Rafa Borrás, PhD^{5,6} and Iñaki Comas^{1,7*}.

- 1 - Institute of Biomedicine of Valencia (CSIC), Valencia (46020), Spain
- 2 - FISABIO Public Health (CSISP), Valencia (46010), Spain
- 3 - Joint Research Unit FISABIO-Universitat de València (CIBERESP), Valencia (46010), Spain
- 4 - La Fe University and Polytechnic Hospital, Valencia, Spain
- 5 - University Clinic Hospital, Valencia (46010), Spain
- 6 - School of Medicine, University of Valencia, Valencia (46010), Spain
- 7 - CIBER in Epidemiology and Public Health, Madrid (28029), Spain

27 **Abstract**

28 Tuberculosis remains one of the main causes of death worldwide. The long and
 29 cumbersome process of culturing *Mycobacterium tuberculosis* complex (MTBC)
 30 bacteria has encouraged the development of specific molecular tools for detecting
 31 the pathogen. Most of these tools aim to become novel tuberculosis diagnostics, and
 32 big efforts and resources are invested in their development, looking for the
 33 endorsement of the main public health agencies. Surprisingly, no study had been
 34 conducted where the vast amount of genomic data available is used to identify the
 35 best MTBC diagnostic markers. In this work, we use large-scale comparative
 36 genomics to provide a catalog of 30 characterized loci that are unique to the MTBC.
 37 Some of these genes could be targeted to assess the physiological status of the
 38 bacilli. Remarkably, none of the conventional MTBC markers is in our catalog. In
 39 addition, we develop a qPCR assay to accurately quantify MTBC DNA in clinical
 40 samples.

41

42

43

44

45

46

47

48

49

50

51

52 Main text

53 Background

54 Tuberculosis (TB) is the most lethal infectious disease caused by a single agent,
 55 namely bacteria belonging to the *Mycobacterium tuberculosis* complex (MTBC)[1].
 56 Whereas isolating the bacteria from clinical specimens is a time-consuming process
 57 that delays both clinical diagnosis and research workflows, rapid molecular tests
 58 have the potential to identify the pathogen DNA in a few hours [2,3]. This is the main
 59 reason why the development of new molecular tools for TB diagnosis is an active
 60 area of research, with many companies involved, looking for the endorsement of the
 61 World Health Organization (WHO) [4]. The most successful example has been the
 62 Xpert MTB/RIF test [5], which was endorsed by the WHO back in 2010 for TB
 63 diagnosis, and recommended as the first-line diagnostic in 2017[6]. Achieving a high
 64 sensitivity and specificity is pivotal for the development and improvement of
 65 molecular tests to ensure an accurate diagnosis. To this end, most tests incorporate
 66 specific markers for the detection of MTBC bacteria. For instance, the new Xpert
 67 MTB/RIF Ultra assay, previously targeting the *rpoB* gene alone, has now
 68 incorporated the insertion sequences IS6110 and IS1081[7]. The insertion sequence
 69 IS6110 has been extensively used as a MTBC-specific marker since first described
 70 in 1990[8]. In addition, the IS6110 can be present in high copy numbers in some
 71 MTBC strains (from 0 to 27 copies)[9], causing the nucleic acid amplification tests
 72 (NAAT) targeting this sequence to achieve higher sensitivities for strains carrying
 73 several copies. However, the specificity of the IS6110 has been questioned since
 74 two decades ago[10–15] what, along with the fact that some strains lack this
 75 insertion sequence, can lead to an incorrect diagnosis[16,17].

76 Several other genes have been used as markers for the accurate identification of
77 MTBC bacteria[18–21]. However, the accuracy of NAATs based on these markers
78 rely on the specificity of the primers, since most of the targeted loci are claimed to be
79 MTBC-specific, yet they were evaluated with limited genomic information on the
80 diversity of NTM and MTBC bacteria.

81 Nowadays, the use of the publicly available *omic* data can help identifying species-
82 specific genetic markers to develop accurate molecular tools. Analyzing *omic* data
83 has been proven to be an effective strategy for the identification of specific markers
84 in several organisms[22–26], and even some workflows have been published for the
85 evaluation of genetic markers based on genomic data[27]. For instance, comparative
86 genomics was used by Zozaya-Valdés *et al.* to assess the population structure of
87 *Mycobacterium chimaera*, identifying six specific loci of these organisms that allowed
88 them to develop a highly accurate qPCR assay.

89 Strikingly, the use of comparative genomics for the identification of MTBC-specific
90 loci has been very limited. The few published studies focused on genetic regions
91 acquired by horizontal gene transfer and used the limited datasets available at the
92 time of publication, a decade ago[28–30]. By contrast, last years have witnessed a
93 burst of available genomic sequences of a wide range of mycobacteria species and
94 thousands of strains of the MTBC[31–33].

95 In this work, we perform a large-scale comparative genomic analysis to provide a
96 reference list of 30 MTBC-specific loci that will be of great utility for the scientific
97 community working on the development of new research and clinical tools for
98 tuberculosis. Remarkably, we found that the main MTBC markers used up to date
99 are also present in other organisms, mainly NTM. In our analysis, we assess the
100 global diversity of each MTBC-specific gene among a comprehensive dataset of

101 more than 4,700 MTBC strains, showing the value of using the genomic data at hand
102 to identify the best targets for diagnostic assays. In addition, we develop a qPCR
103 assay based on one of these markers capable of quantifying MTBC DNA in clinical
104 samples.

105

106 **Methods**

107 **In silico identification of MTBC-specific diagnostic gene markers**

108 To identify MTBC-specific loci, we used blastn[34] to look for all the genes of the
109 tuberculosis reference strain H37Rv (NC_000962.3) in the NCBI nucleotide non-
110 redundant database (accessed October 2018) and a custom database comprising
111 4,277 NTM assemblies (Supplementary Methods 1). All the searches were
112 performed specifying the algorithm blastn with a word size (or seed) of 7 bp. Then,
113 we filtered the results with a set of stringent parameters to discard loci similar to any
114 genomic region of any organism other than MTBC. We discarded all the genes that
115 presented an alignment of more than 25% of its sequence (query coverage) with a
116 similarity greater than 80%. If a gene was aligned in 60% of its sequence or longer it
117 was discarded regardless of the similarity of the alignment. We only kept those
118 genes that were present in all the MTBC bacteria.

119 Once potential MTBC-specific markers were identified, we decided to assess their
120 genetic diversity. To do this, we analyzed the polymorphisms (single nucleotide
121 polymorphisms (SNPs) and indels) observed at each position across a dataset
122 comprising 4,766 genomes of MTBC strains[35]. Therefore, the number of SNPs of
123 each gene was calculated as the sum of positions showing any nucleotide other than
124 the reference. In the case of indels, we considered those positions showing an indel
125 in at least 10 strains (0.2% of the database) to avoid the noise introduced by single-

126 strain indels spanning large genic regions and possible false deletions arising as a
127 result of sequencings with uneven genomic coverages. This allowed us to calculate
128 different metrics for each gene such as the absolute number of polymorphisms,
129 polymorphisms per base and, most importantly, the prevalence of each one.
130 Finally, we looked for available information of these genes in the bibliography, what
131 allowed us to discard some candidates based on their genomic context and provide
132 extended information about their physiology. We gathered transcriptomic and
133 proteomic data derived from different published studies: transcriptomic data in
134 response to overexpression of 206 transcription factors[36], different genotoxic
135 stresses[37] and response to nitric oxide stress at different time-points[38], as well
136 as proteomic data in response to nutrient starvation[39].

137

138 **Set-up of a MTBC-specific qPCR assay for DNA detection and quantification**

139 We used the list of 30 MTBC-specific loci to set up a qPCR assay for the detection
140 and quantification of MTBC DNA. To select the target for the assay, we took into
141 consideration the number of polymorphisms per base, the absence of high-prevalent
142 polymorphisms, the gene length and its genomic context. These criteria enabled an
143 optimum design of primers, amplifying a universal and highly-specific region for the
144 detection of MTBC. We designed the primers and probes for the assay using the
145 web tool Primer-BLAST[40], checking that no unspecific amplicons were predicted.
146 Finally, the qPCR assay consisted on the amplification of a 65 bp region within the
147 Rv2341 gene using the following primers: Forward-GCCGCTCATGCTCCTTGGAT,
148 Reverse-AGGTCGGTTCGCTGGTCTTG, Probe-
149 TGAGTGCCTGCGGCCGCAGCGC.

150 To test the specificity of the assay we performed qPCR experiments with DNA from
 151 all MTBC lineages (except lineage 7 due to unavailability), human DNA, a mock
 152 sample with mixed DNA from 20 different bacterial species (ATCC[®] MSA-1002[™])
 153 and 17 different species of NTM (Supplementary Methods 2).
 154 The reaction efficiency was calculated using serial dilutions of pure H37Rv DNA as
 155 template (0.5 ng/ul to 0.5×10^{-5} ng/ul). In addition, we evaluated the performance of
 156 the assay detecting and quantifying MTBC DNA in a test set of clinical samples. We
 157 used extracted DNA from 12 homogenized sputum samples from culture-positive TB
 158 patients, two of them with negative smear microscopy. We also used a DNA
 159 extraction from a non-TB patient sputum to spike in known concentrations of pure
 160 H37Rv DNA (0.5 ng/ul to 0.5×10^{-5} ng/ul), to calculate the reaction efficiency in
 161 clinical samples.
 162 All the qPCR reactions were carried out using hydrolysis probes chemistry
 163 (FAM/BHQ) in a total volume of 20ul, containing 10ul of Kapa Probe Fast Master Mix
 164 2X (Kapa Biosystems), 250mM of each primer, 350mM of probe and 2ul of sample.
 165 All were performed in a Roche Lightcycler 96 (Roche Diagnostics), with two
 166 replicates per sample and including reactions with no template as negative controls
 167 (NTC). When calculating reaction efficiencies, we used three replicates per point
 168 instead of two. The conditions for each assay comprised an initial denaturation step
 169 at 95°C for 3 minutes, followed by 55 amplification cycles as follows: 20 seconds at
 170 60°C for annealing, 1 second at 72°C for extension, and 10 seconds at 95°C for
 171 denaturation. The results were analyzed with LightCycler 96[®] 1.1 software.
 172 Triplicates of each assay were carried out to check the reproducibility.

173

174 **Bacterial culture, clinical specimens and DNA extraction.**

175 All the DNA extractions were performed in our laboratory except for the commercial
176 DNA mix of 20 bacterial species. Available cultures of different NTM species were
177 subcultured in in 7H11 solid agar media and then the DNA extracted following the
178 standard CTAB protocol[41] with an inactivation step of 1 hour at 80°C. DNA
179 concentrations were measured with the Qubit fluorometer (dsDNA high-sensitivity
180 kit) and samples with a concentration higher than 1ng/ul were normalized to 1ng/ul.
181 In the case of the 13 sputum specimens, DNA extraction was performed as
182 described by Votintseva *et al*[42]. All the samples were handled in a BSL-3 until DNA
183 was extracted and purified.

184 **Ethics approval**

185 The clinical specimens used in this study were collected as part of the surveillance
186 program of communicable diseases by the General Directorate of Public Health of
187 the Comunidad Valenciana and, as such, falls outside the mandate of the
188 corresponding Ethics Committee for Biomedical Research. All personal information
189 was anonymized and no data allowing individual identification was retained.

190

191 **Results**

192

193 We identified 40 genes to be uniquely present in members of the MTBC according to
194 our filtering parameters (Figure 1). After evaluating their genetic diversity across a
195 database of more than 4,700 MTBC strains, we observed that the median number of
196 SNPs per base was 0.07, with some of these genes showing either higher or lower
197 diversities (up to 0.1 and 0.04 SNPs/base respectively), probably as a result of

198 different selective pressures. Importantly, although most of the polymorphisms
199 analyzed were strain-specific, we observed high prevalent polymorphisms as well
200 (Figure 1, Supplementary File 1). For instance, Rv0610c showed a SNP present in
201 4182 strains and Rv2823c showed an insertion in 4,345 strains. Analysis of the
202 phylogenetic distribution of these polymorphisms confirmed that they mapped to
203 deep branches in the phylogeny. For example, the SNP in Rv0610c affected all
204 modern lineages (L2, L3, L4).

205 Among these, 9 genes were discarded as potential diagnostic markers since they
206 were included in regions of difference (RD) 182 (Rv2274c) and RD 207 (Rv2816c-
207 Rv2820c) as described in Gagneux *et al.*[43] or were in variable genomic regions
208 associated to CRISPR elements (Rv2816c-2823c)[44]. Another gene, Rv3424c was
209 also discarded as we found it to be duplicated in a very labile genomic region,
210 between the (putative) transposase of the insertion sequence IS1532 and PPE 59.
211 Therefore, the curated list of MTBC-specific diagnostic markers finally consisted in
212 30 genes (Figure 1).

213 When looking at published transcriptomic and proteomic data (see Methods), we
214 observed that Rv2003c, Rv2142c, and Rv3472 proteins are found in greater levels
215 (6.19, 3.6 and 100-fold respectively) when the bacteria is subjected to starvation.
216 Interestingly, Rv2003c is also observed to be overexpressed upon treatment with
217 nitric oxide (Supplementary File 2).

218 Based on our large genomic analysis, we set up a qPCR assay targeting the Rv2341
219 gene. This gene, described as “probable conserved lipoprotein lppQ” in the
220 Mycobrowser database[45], is situated in a stable genomic region, between the
221 asparagine tRNA and the gene of the DNA primase, involved in the synthesizes of
222 the okazaki fragments. Furthermore, we were able to design an optimized set of

primers that avoid, at the same time, any region harboring prevalent polymorphisms (Figure 1).

When testing the qPCR assay with a panel of samples including different MTBC lineages, human, mock bacterial communities and different NTMs, the specificity of the assay was of 100%. The efficiency of the reaction was of 95% showing a limit of detection of 10fg (hypothetically corresponding to 2 genome equivalents). When using a standard curve of pure H37Rv DNA spiked in sputum samples, both the efficiency of the reaction (97%) and the limit of detection remained unaltered (Figure 2). When testing our qPCR assay with a panel of 12 TB sputum samples, we were able to detect and quantify MTBC DNA in all TB patient sputa, including 2 confirmed TB cases with a negative smear microscopy (Supplementary File 4).

Discussion

Identification of MTBC markers for the development of new diagnostic and research tools for tuberculosis has been an active area of research over the last decades, focusing on the direct or indirect detection of the tubercle bacilli. It is striking that for such a relevant disease, from both the epidemiological and economical point of view, for which tons of genomic data is already available, the identification of MTBC-specific genes had been relegated to the background. This has been probably motivated by the fact that current molecular tools have shown to perform well in most of situations. For instance, assays targeting the insertion sequence IS6110 ([46] or *rpoB*[47]. However, the available tools are not enough to stop the spread of the disease and for this reason many new generation diagnostics are still being developed with the aim to improve the accuracy of the existing ones and tackle their known flaws.

248 Our analysis provides invaluable information to develop such diagnostics, with a
249 catalog of specific MTBC markers. Remarkably, some of the markers that we identify
250 could be targeted to determine the physiological status of MTBC bacteria under
251 certain conditions. For example, Rv2003c, overexpressed during starvation and
252 upon treatment with nitric oxide[38,39], is also upregulated during dormancy[48].
253 Similarly, Rv1374c has been described to be a small RNA that is highly expressed
254 during exponential growth[49], and hence could be used to evaluate the replicative
255 state of the bacilli.

256 Strikingly, none of the markers considered to be MTBC-specific up to date are in our
257 list of unique MTBC genes. For instance, when examining in which species the
258 IS6110 can be found, we observed several non-MTBC organisms, including 14
259 NTMs, carrying at least one copy. The same is true for IS1081 and *mpt64*, present in
260 38 and 6 NTM respectively (Supplementary File 3). Similarly, the short-chain
261 dehydrogenase/reductase gene (SDR) (Rv0303, region 365,234–366,142), which
262 has been recently described as a *M. tuberculosis*-specific marker[28], is actually
263 present in several NTM, as revealed by a blastn search in the non-redundant
264 database of the NCBI web server (accessed January 2019), and in our database of
265 NTM assemblies (Supplementary File 5). The fact that IS6110 is still one of the most
266 used genetic targets for MTBC DNA detection (for example in the new Xpert Ultra
267 MTB/RIF assay[7]), highlights the great utility, and the necessity, of translating the
268 results of genomic analyses to the laboratory.

269 To illustrate the translational potential of our work, we set up an accurate qPCR
270 assay capable of quantifying MTBC DNA with 100% specificity and a sensitivity up to
271 2 genome copies. Quantifying MTBC DNA from clinical samples is challenging due
272 to the presence of PCR inhibitors along with great proportions of DNA from human

273 and oropharyngeal microbiota. However, this capability is invaluable not only for
274 diagnostic purposes, but also in the research context, for example when developing
275 new protocols[42,50]. Remarkably, our assay, targeting a small region of the Rv2341
276 gene, showed an excellent performance in a test set of clinical specimens. However,
277 we want to highlight that the list provided here comprehends 30 loci, from which
278 many different molecular tools for tuberculosis could be developed.

279

280 Altogether, our analysis has a direct translational value, as it represents an important
281 resource for research groups and companies involved in the development and
282 improvement of novel TB diagnostics. For instance, the markers identified in this
283 work could be used to improve existing tests such as the Xpert MTB/RIF assay, by
284 including targets that we have demonstrated to be globally conserved and fully
285 specific to the MTBC.

286

287 **Declarations**

288 **Competing Interests**

289 The authors declare no conflict of interest in this article.

290

291 **Funding Sources**

292 This work was supported by projects of the European Research Council (ERC)
293 (638553-TB-ACCELERATE), Ministerio de Economía y Competitividad, and
294 Ministerio de Ciencia, Innovación y Universidades (Spanish Government), SAF2013-
295 43521-R, SAF2016-77346-R and SAF2017-92345-EXP (to IC), BES-2014-071066
296 (to GAG), FPU 13/00913 (to ACO)

297

298 **Author Contributions**

299 GAG and IC designed the study and analyzed the data. GAG and ACO analyzed the
300 4,766 MTBC strains dataset. GAG, MTP and CML performed the qPCR
301 experiments. GAG and LMV cultured the non-tuberculous mycobacteria and
302 performed the DNA extractions. AGB and RB did the microbiological identification of
303 the isolates of non-tuberculous mycobacteria. RB provided the clinical specimens
304 and clinical data. GAG and IC wrote the first draft of the manuscript. All authors
305 contributed to the final version of the manuscript.

306

307 **Contact Information**

308 Corresponding author e-mail, Iñaki Comas: icomas@ibv.csic.es

309

310 **References**

- 311 1. World Health Organization. Global Tuberculosis Report 2017. 2017.
- 312 2. Eddabra R, Benhassou HA. Rapid molecular assays for detection of
313 tuberculosis. *Pneumonia. BioMed Central*. **2018**; 10(1):4.
- 314 3. Machado D, Couto I, Viveiros M. Advances in the molecular diagnosis of
315 tuberculosis: From probes to genomes. *Infect Genet Evol*. **2018**; Available from:
316 <http://dx.doi.org/10.1016/j.meegid.2018.11.021>
- 317 4. Pai M, Nicol MP, Boehme CC. Tuberculosis Diagnostics: State of the Art and
318 Future Directions. *Microbiol Spectr*. **2016**; 4(5). Available from: [http://dx.doi.org/](http://dx.doi.org/10.1128/microbiolspec.TBTB2-0019-2016)
319 [10.1128/microbiolspec.TBTB2-0019-2016](http://dx.doi.org/10.1128/microbiolspec.TBTB2-0019-2016)

- 320 5. Cirillo DM, Miotto P, Tortoli E. Evolution of Phenotypic and Molecular Drug
321 Susceptibility Testing. *Adv Exp Med Biol.* **2017**; 1019:221–246.
- 322 6. WHO meeting report of a technical expert consultation: non-inferiority analysis
323 of Xpert MTF/RIF Ultra compared to Xpert MTB/RIF. Geneva: World Health
324 Organization; **2017** (WHO/HTM/TB/2017.04); Available from: <http://www.who.int/tb/publications/2017/XpertUltra/en/>
325
- 326 7. Dorman SE, Schumacher SG, Alland D, et al. Xpert MTB/RIF Ultra for detection
327 of *Mycobacterium tuberculosis* and rifampicin resistance: a prospective
328 multicentre diagnostic accuracy study. *Lancet Infect Dis.* **2018**; 18(1):76–84.
- 329 8. Thierry D, Brisson-Noël A, Vincent-Lévy-Frébault V, Nguyen S, Guesdon JL,
330 Gicquel B. Characterization of a *Mycobacterium tuberculosis* insertion
331 sequence, IS6110, and its application in diagnosis. *J Clin Microbiol.* **1990**;
332 28(12):2668–2673.
- 333 9. Roychowdhury T, Mandal S, Bhattacharya A. Analysis of IS6110 insertion sites
334 provide a glimpse into genome evolution of *Mycobacterium tuberculosis*. *Sci*
335 *Rep.* **2015**; 5:12567.
- 336 10. Kent L, McHugh TD, Billington O, Dale JW, Gillespie SH. Demonstration of
337 homology between IS6110 of *Mycobacterium tuberculosis* and DNAs of other
338 *Mycobacterium spp.* *J Clin Microbiol.* **1995**; 33(9):2290–2293.
- 339 11. Liébana E, Aranaz A, Francis B, Cousins D. Assessment of genetic markers for
340 species differentiation within the *Mycobacterium tuberculosis* complex. *J Clin*
341 *Microbiol.* **1996**; 34(4):933–938.

- 342 12. McHugh TD, Newport LE, Gillespie SH. IS6110 homologs are present in
343 multiple copies in mycobacteria other than tuberculosis-causing mycobacteria. J
344 Clin Microbiol. **1997**; 35(7):1769–1771.
- 345 13. Hellyer TJ, DesJardin LE, Beggs ML, et al. IS6110 homologs are present in
346 multiple copies in mycobacteria other than tuberculosis-causing mycobacteria. J
347 Clin Microbiol. **1998**; 36(3):853–854.
- 348 14. Müller R, Roberts CA, Brown TA. Complications in the study of ancient
349 tuberculosis: non-specificity of IS6110 PCRs. STAR: Science & Technology of
350 Archaeological Research. **2015**; 1(1):1–8.
- 351 15. Coros A, DeConno E, Derbyshire KM. IS6110, a *Mycobacterium tuberculosis*
352 complex-specific insertion sequence, is also present in the genome of
353 *Mycobacterium smegmatis*, suggestive of lateral gene transfer among
354 mycobacterial species. J Bacteriol. **2008**; 190(9):3408–3410.
- 355 16. Steensels D, Fauville-Dufaux M, Boie J, De Beenhouwer H. Failure of PCR-
356 Based IS6110 analysis to detect vertebral spondylodiscitis caused by
357 *Mycobacterium bovis*. J Clin Microbiol. **2013**; 51(1):366–368.
- 358 17. Huyen MNT, Tiemersma EW, Kremer K, et al. Characterisation of
359 *Mycobacterium tuberculosis* isolates lacking IS6110 in Viet Nam. Int J Tuberc
360 Lung Dis. **2013**; 17(11):1479–1485.
- 361 18. Therese KL, Jayanthi U, Madhavan HN. Application of nested polymerase chain
362 reaction (nPCR) using MPB 64 gene primers to detect *Mycobacterium*
363 *tuberculosis* DNA in clinical specimens from extrapulmonary tuberculosis
364 patients. Indian J Med Res. **2005**; 122(2):165–170.

- 365 19. Chakravorty S, Sen MK, Tyagi JS. Diagnosis of extrapulmonary tuberculosis by
366 smear, culture, and PCR using universal sample processing technology. J Clin
367 Microbiol. **2005**; 43(9):4357–4362.
- 368 20. Nimesh M, Joon D, Pathak AK, Saluja D. Comparative study of diagnostic
369 accuracy of established PCR assays and in-house developed sdaA PCR
370 method for detection of *Mycobacterium tuberculosis* in symptomatic patients
371 with pulmonary tuberculosis. J Infect. **2013**; 67(5):399–407.
- 372 21. Queipo-Ortuño MI, Colmenero JD, Bermudez P, Bravo MJ, Morata P. Rapid
373 differential diagnosis between extrapulmonary tuberculosis and focal
374 complications of brucellosis using a multiplex real-time PCR assay. PLoS One.
375 **2009**; 4(2):e4526.
- 376 22. Carmona SJ, Sartor PA, Leguizamón MS, Campetella OE, Agüero F. Diagnostic
377 peptide discovery: prioritization of pathogen diagnostic markers using multiple
378 features. PLoS One. **2012**; 7(12):e50748.
- 379 23. Buchanan CJ, Webb AL, Mutschall SK, et al. A Genome-Wide Association
380 Study to Identify Diagnostic Markers for Human Pathogenic *Campylobacter*
381 *jejuni* Strains. Front Microbiol. **2017**; 8:1224.
- 382 24. Carrera M, Böhme K, Gallardo JM, Barros-Velázquez J, Cañas B, Calo-Mata P.
383 Characterization of Foodborne Strains of *Staphylococcus aureus* by Shotgun
384 Proteomics: Functional Networks, Virulence Factors and Species-Specific
385 Peptide Biomarkers. Front Microbiol. **2017**; 8:2458.
- 386 25. Wang H, Drake SK, Yong C, et al. A Genoproteomic Approach to Detect
387 Peptide Markers of Bacterial Respiratory Pathogens. Clin Chem. **2017**;

- 388 63(8):1398–1408.
- 389 26. Koul S, Kumar P. A Unique Genome Wide Approach to Search Novel Markers
390 for Rapid Identification of Bacterial Pathogens. J Mol Genet Med. **2015**; 09(04).
- 391 27. Felten A, Guillier L, Radomski N, Mistou M-Y, Lailier R, Cadel-Six S. Genome
392 Target Evaluator (GTEvaluator): A workflow exploiting genome dataset to
393 measure the sensitivity and specificity of genetic markers. PLoS One. **2017**;
394 12(7):e0182082.
- 395 28. Kakhki RK, Neshani A, Sankian M, Ghazvini K, Hooshyar A, Sayadi M. The
396 short-chain dehydrogenases/reductases (SDR) gene: A new specific target for
397 rapid detection of *Mycobacterium tuberculosis* complex by modified comparative
398 genomic analysis. Infect Genet Evol. **2019**; Available from:
399 <http://dx.doi.org/10.1016/j.meegid.2019.01.012>
- 400 29. Becq J, Gutierrez MC, Rosas-Magallanes V, et al. Contribution of horizontally
401 acquired genomic islands to the evolution of the tubercle bacilli. Mol Biol Evol.
402 **2007**; 24(8):1861–1871.
- 403 30. Veyrier F, Pletzer D, Turenne C, Behr MA. Phylogenetic detection of horizontal
404 gene transfer during the step-wise genesis of *Mycobacterium tuberculosis*. BMC
405 Evol Biol. **2009**; 9:196.
- 406 31. Fedrizzi T, Meehan CJ, Grottola A, et al. Genomic characterization of
407 Nontuberculous Mycobacteria. Sci Rep. Nature Publishing Group; **2017**;
408 7:45258.
- 409 32. Coll F, Phelan J, Hill-Cawthorne GA, et al. Genome-wide analysis of multi- and

- 410 extensively drug-resistant *Mycobacterium tuberculosis*. Nat Genet. Nature
411 Publishing Group; **2018**; 50(2):307.
- 412 33. CRyPTIC Consortium and the 100,000 Genomes Project, Allix-Béguec C,
413 Arandjelovic I, et al. Prediction of Susceptibility to First-Line Tuberculosis Drugs
414 by DNA Sequencing. N Engl J Med. **2018**; 379(15):1403–1415.
- 415 34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment
416 search tool. J Mol Biol. **1990**; 215(3):403–410.
- 417 35. Chiner-Oms Á, González-Candelas F, Comas I. Gene expression models based
418 on a reference laboratory strain are poor predictors of *Mycobacterium*
419 *tuberculosis* complex transcriptional diversity. Sci Rep. **2018**; 8(1):3813.
- 420 36. Turkarslan S, Peterson EJR, Rustad TR, et al. A comprehensive map of
421 genome-wide gene regulation in *Mycobacterium tuberculosis*. Sci Data. **2015**;
422 2:150010.
- 423 37. Namouchi A, Gómez-Muñoz M, Frye SA, et al. The *Mycobacterium tuberculosis*
424 transcriptional landscape under genotoxic stress. BMC Genomics. **2016**;
425 17(1):791.
- 426 38. Cortes T, Schubert OT, Banaei-Esfahani A, Collins BC, Aebersold R, Young DB.
427 Delayed effects of transcriptional responses in *Mycobacterium tuberculosis*
428 exposed to nitric oxide suggest other mechanisms involved in survival. Sci Rep.
429 **2017**; 7(1):8208.
- 430 39. Albrethsen J, Agner J, Piersma SR, et al. Proteomic profiling of *Mycobacterium*
431 *tuberculosis* identifies nutrient-starvation-responsive toxin-antitoxin systems. Mol

- 432 Cell Proteomics. **2013**; 12(5):1180–1191.
- 433 40. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. Primer-
434 BLAST: a tool to design target-specific primers for polymerase chain reaction.
435 BMC Bioinformatics. **2012**; 13:134.
- 436 41. Somerville W, Thibert L, Schwartzman K, Behr MA. Extraction of
437 *Mycobacterium tuberculosis* DNA: a question of containment. J Clin Microbiol.
438 **2005**; 43(6):2996–2997.
- 439 42. Votintseva AA, Bradley P, Pankhurst L, et al. Same-day diagnostic and
440 surveillance data for tuberculosis via whole genome sequencing of direct
441 respiratory samples [Internet]. J Clin Microbiol. **2017**; 55(5):1285–1298.
- 442 43. Gagneux S, DeRiemer K, Van T, et al. Variable host-pathogen compatibility in
443 *Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A. **2006**; 103(8):2869–
444 2873.
- 445 44. Freidlin PJ, Nissan I, Luria A, et al. Structure and variation of CRISPR and
446 CRISPR-flanking regions in deleted-direct repeat region *Mycobacterium*
447 *tuberculosis* complex strains. BMC Genomics. **2017**; 18(1):168.
- 448 45. Kapopoulou A, Lew JM, Cole ST. The MycoBrowser portal: A comprehensive
449 and manually annotated resource for mycobacterial genomes. Tuberculosis .
450 **2011**; 91(1):8–13.
- 451 46. Harkins KM, Buikstra JE, Campbell T, et al. Screening ancient tuberculosis with
452 qPCR: challenges and opportunities. Philos Trans R Soc Lond B Biol Sci. **2015**;
453 370(1660):20130622.

47. Li S, Liu B, Peng M, et al. Diagnostic accuracy of Xpert MTB/RIF for tuberculosis detection in different regions with different endemic burden: A systematic review and meta-analysis. PLoS One. Public Library of Science; **2017**; 12(7):e0180725.
48. Hegde SR, Rajasingh H, Das C, Mande SS, Mande SC. Understanding communication signals during mycobacterial latency through predicted genome-wide protein interactions and boolean modeling. PLoS One. **2012**; 7(3):e33893.
49. Arnvig KB, Comas I, Thomson NR, et al. Sequence-Based Analysis Uncovers an Abundance of Non-Coding RNA in the Total Transcriptome of *Mycobacterium tuberculosis*. PLoS Pathog. Public Library of Science; **2011**; 7(11):e1002342.
50. Brown AC, Bryant JM, Einer-Jensen K, et al. Rapid Whole-Genome Sequencing of *Mycobacterium tuberculosis* Isolates Directly from Clinical Samples. J Clin Microbiol. **2015**; 53(7):2230–2237.

Figure Legends

Figure 1. The 40 *Mycobacterium tuberculosis* complex (MTBC)-specific loci identified after an extensive search with blast in the NCBI non-redundant nucleotide database and a custom database of 4,277 Non-tuberculous mycobacteria (NTM). Gene names in red indicate loci that were discarded as diagnostic markers for being within regions of difference (Rv2274c within RD 182 and Rv2816c-2820c within RD 207), associated to CRISPR (Rv2816c-2823c) or duplicated in the genome (Rv3424c). Concentric circles represent genetic diversity metrics calculated by analyzing a dataset of 4,766 MTBC strains. Outer circle: heatmap representing the

478 number of SNPs per base. Blue circle: prevalence of each SNP of each gene across
 479 the database of MTBC strains. Inner, read circle: prevalence of each indel of each
 480 gene across the database of MTBC strains. Note that both inner circles have two
 481 scales, one from 0 to 300 strains and other from 300 to 4,800 strains. The region of
 482 the Rv2341 gene amplified in our qPCR assay, avoiding prevalent polymorphisms, is
 483 indicated in light yellow. Note that regions of difference 182 and 207 are clearly
 484 detected in our analysis, indicated as contiguous deleted regions in a high number of
 485 strains.

486

487 **Figure 2.** Standard curve for the qPCR assay targeting Rv2341 using known
 488 quantities of pure H37Rv DNA (in blue; efficiency=95%) and pure H37Rv DNA
 489 spiked in sputum samples (in red; efficiency=97%). In the upper x-axis is
 490 represented the hypothetical number of genome copies. All qPCR experiments were
 491 carried out in triplicates to check for reproducibility.

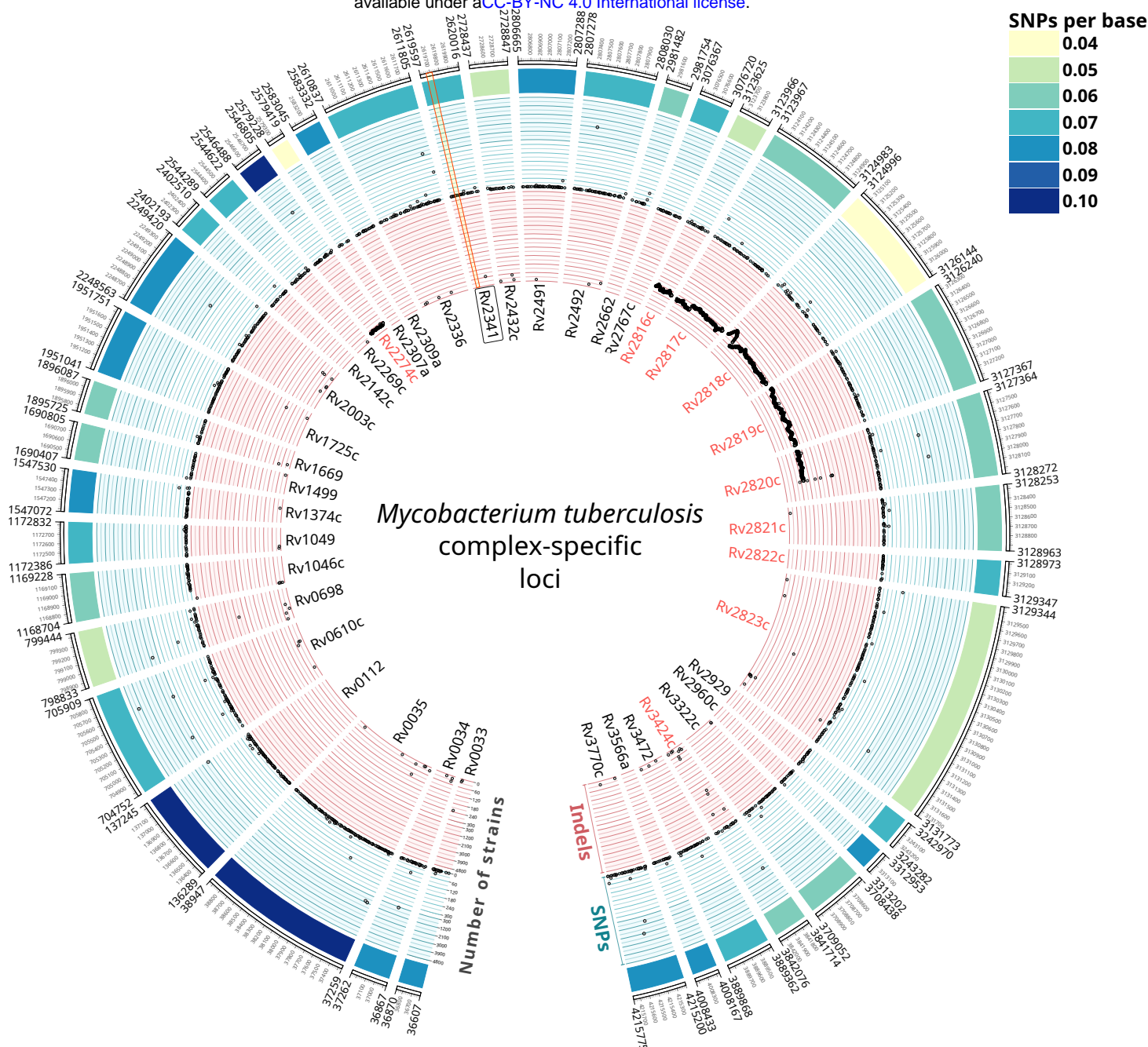


Figure 1. The 40 *Mycobacterium tuberculosis* complex (MTBC)-specific loci identified after an extensive search with blast in the NCBI non-redundant nucleotide database and a custom database of 4,277 Non-tuberculous mycobacteria (NTM). Gene names in red indicate loci that were discarded as diagnostic markers for being within regions of difference (Rv2274c within RD 182 and Rv2816c-2820c within RD 207), associated to CRISPR (Rv2816c-2823c) or duplicated in the genome (Rv3424c). Concentric circles represent genetic diversity metrics calculated by analyzing a dataset of 4,766 MTBC strains. Outer circle: heatmap representing the number of SNPs per base. Blue circle: prevalence of each SNP of each gene across the database of MTBC strains. Inner, read circle: prevalence of each indel of each gene across the database of MTBC strains. Note that both inner circles have two scales, one from 0 to 300 strains and other from 300 to 4,800 strains. The region of the Rv2341 gene amplified in our qPCR assay, avoiding prevalent polymorphisms, is indicated in light yellow. Note that regions of difference 182 and 207 are clearly detected in our analysis, indicated as contiguous deleted regions in a high number of strains.

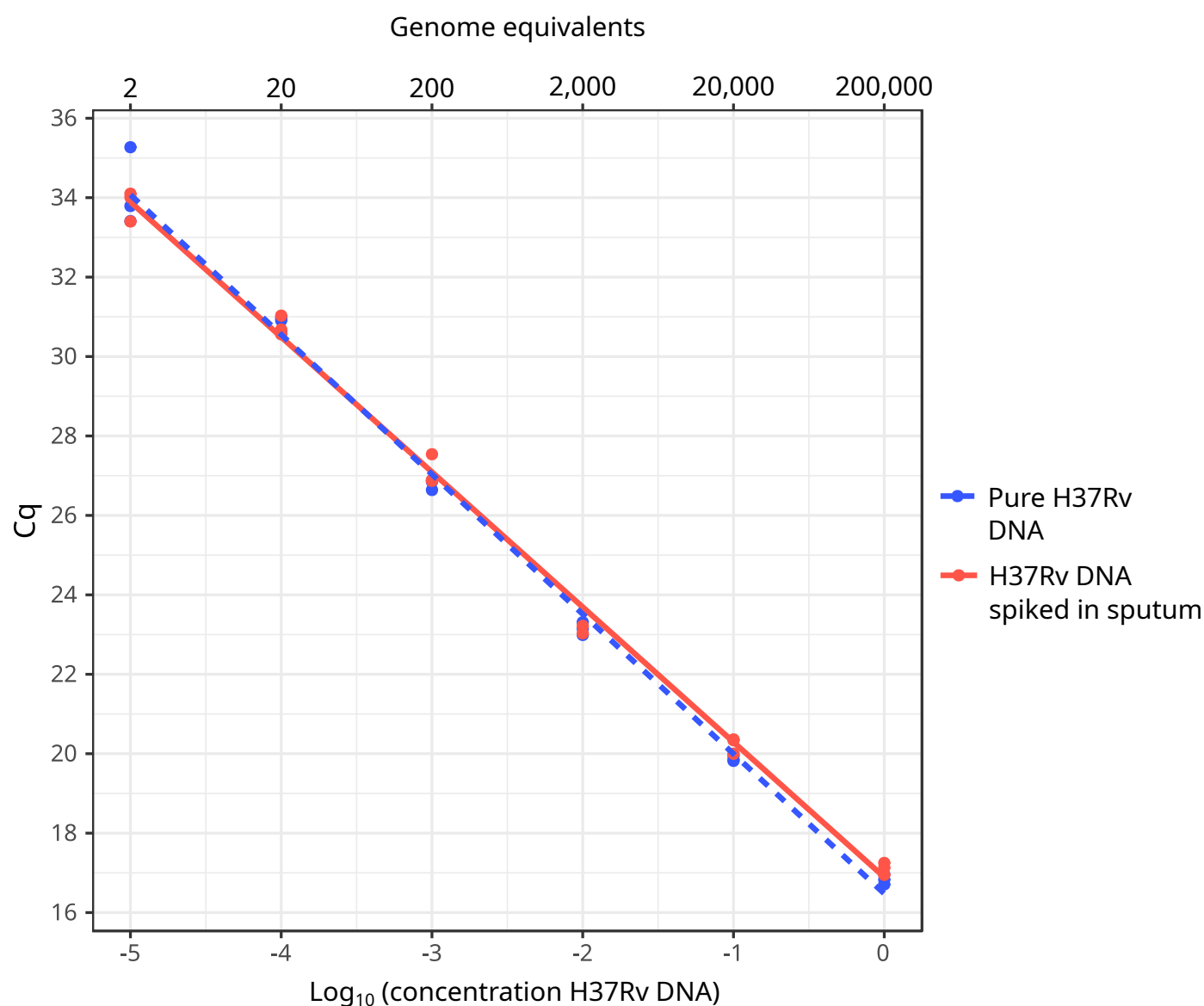


Figure 2. Standard curve for the qPCR assay targeting Rv2341 using known quantities of pure H37Rv DNA (in blue; efficiency=95%) and pure H37Rv DNA spiked in sputum samples (in red; efficiency=97%). In the upper x-axis is represented the hypothetical number of genome copies. All qPCR experiments were carried out in triplicates to check for reproducibility.