Genetic signatures of *Mycobacterium tuberculosis* Nonthaburi genotype revealed by whole genome analysis of isolates from tuberculous meningitis patients in Thailand

Olabisi Oluwabukola Coker, Angkana Chaiprasert, Chumpol Ngamphiw, Sissades Tongsima, Sanjib Mani Regmi, Taane G Clark, Rick Twee Hee Ong, Yik-Ying Teo, Therdsak Prammananan, Prasit Palittapongarnpim

Genome sequencing plays a key role in understanding the genetic diversity of *Mycobacterium tuberculosis (M.tb)*. The genotype-specific character of *M. tb* contributes to tuberculosis severity and emergence of drug resistance. Strains of *M. tb* complex can be classified into seven lineages. The Nonthaburi (NB) genotype, belonging to the Indo-Oceanic lineage (lineage 1), has a unique spoligotype and IS*6110*-RFLP pattern but has not previously undergone a detailed whole genome analysis. In addition, there is not much information available on the whole genome analysis of *M. tb* isolates from tuberculous meningitis (TBM) patients in public databases. Isolates CSF3053, 46-5069 and 43-13838 of NB genotype were obtained from the cerebrospinal fluids of TBM Thai patients in Siriraj Hospital, Bangkok. The whole genomes were subjected to high throughput sequencing. The sequence data of each isolate were assembled into draft genome. The sequences were also aligned to reference genome, to determine genomic variations. Single nucleotide polymorphisms (SNPs) were obtained and grouped according to the functions of the genes containing them. They were compared with SNPs from 1,601 genomes, representing the seven lineages of *M. tb* complex, to determine the uniqueness of NB genotype. Susceptibility to first-line, second-line and other antituberculosis drugs were determined and related to the SNPs previously reported in drug resistant related genes. The assembled genomes have an average size of 4,364,461 bp, 4,154 genes, 48 RNAs and 64 pseudogenes. A 500 base pairs deletion, which includes *ppe50*, was found in all isolates. RD239, specific for members of Indo Oceanic lineage, and RD147c were identified. A total of 2,202 SNPs were common to the isolates and used to classify the NB strains as members of sublineage 1.2.1. Compared with 1,601 genomes from the seven lineages of *M. tb* complex, mutation G2342203C was found novel to the isolates in this study. Three mutations (T28910C, C1180580T and C152178T) were found only in Thai NB isolates, including isolates from previous study. Although drug susceptibility tests indicated pan-susceptibility, non-synonymous SNPs previously reported to be associated with resistance to anti-tuberculous drugs; isoniazid, ethambutol, and ethionamide were

identified in all the isolates. Non-synonymous SNPs were found in virulence genes such as the genes playing roles in apoptosis inhibition and phagosome arrest. We also report polymorphisms in essential genes, efflux pumps associated genes and genes with known epitopes. The analysis of the TBM isolates and the availability of the variations obtained will provide additional resources for global comparison of isolates from pulmonary tuberculosis and TBM. It will also contribute to the richness of genomic databases towards the prediction of antibiotic resistance, level of virulence and of origin of infection.

1   **Title page**

3   Genetic signatures of *Mycobacterium tuberculosis* Nonthaburi genotype revealed by whole

4   genome analysis of isolates from tuberculous meningitis patients in Thailand.

6   Olabisi Oluwabukola Coker[1], Angkana Chaiprasert[1#], Chumpol Ngamphiw[2], Sissades

7   Tongsima[2], Sanjib Mani Regmi[3], Taane G. Clark[4,5], Ong Twee Hee[6], Teo Yik Ying[6] Therdsak

8   Prammananan[2], Prasit Palittapongarnpim[7],

10 [1]Department of Microbiology, Faculty of Medicine Siriraj Hospital, Mahidol University,

11 Bangkok 10700, Thailand, bisistill@yahoo.com

12 [2]National Center for Genetic Engineering and Biotechnology, National Science and Technology

13 Development Agency, Pathum Thani 12120, Thailand.

14 [3] Department of Microbiology, Gandaki Medical Collage, Pokhara, Kaski, Nepal.

15 [4] Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine,

16 London, United Kingdom

17 [5] Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical

18 Medicine, London, United Kingdom.

19 [6] Saw Swee Hock School of Public Health, National University of Singapore, Singapore.

20 [7] Department of Microbiology, Faculty of Science Mahidol University, Bangkok 10400,

21 Thailand.

25   #Corresponding author

26   Angkana Chaiprasert, Dr. rer. nat.

27   Department of Microbiology

28   Faculty of Medicine Siriraj Hospital, Mahidol University

29   2 Wanglang Rd., Bangkok-Noi, Bangkok 10700, Thailand

30   Tel. 66-2-419-8256

31   Fax. 66-2-418-2054

32   E-mail. angkana.cha@mahidol.ac.th

40

41

42

43

44

45

46

47

48

49

50

51

52   **Abstract**

53   **Background.** Genome sequencing plays a key role in understanding the genetic diversity of

54   *Mycobacterium tuberculosis*. The genotype-specific character of *M. tuberculosis* contributes to

55   tuberculosis severity and emergence of drug resistance. Strains of *M. tuberculosis* complex can

56   be classified into seven lineages. The Nonthaburi genotype, belonging to the Indo-Oceanic

57   lineage (lineage 1), has a unique spoligotype and IS*6110*-RFLP pattern but has not previously

58   undergone a detailed whole genome analysis. In addition, there is not much information

59   available on the whole genome analysis of *M. tuberculosis* isolates from tuberculous meningitis

60   patients in public databases.

61   **Methods**. Isolates CSF3053, 46-5069 and 43-13838 of Nonthaburi genotype were obtained from

62   the cerebrospinal fluids of tuberculous meningitis Thai patients in Siriraj Hospital, Bangkok. The

63   whole genomes were subjected to high throughput   sequencing. The sequence data of each

64   isolate were assembled into draft genome.

65   The sequences were also aligned to reference genome, to determine genomic variations. Single

66   nucleotide polymorphisms (SNPs) were obtained and grouped according to the functions of the

67   genes containing them. They were compared with SNPs from 1,601 genomes, representing the

68   seven lineages of *M. tuberculosis* complex, to determine the uniqueness of Nonthaburi genotype.

69   Susceptibility to first-line, second-line and other antituberculosis drugs were determined and

70   related to the SNPs previously reported in drug resistant related genes.

71   **Results.** The assembled genomes have an average size of 4,364,461 bp, 4,154 genes, 48 RNAs

72   and 64 pseudogenes. A 500 base pairs deletion, which includes *ppe50,* was found in all isolates.

73   RD239, specific for members of Indo Oceanic lineage, and RD147c were identified.

74   A total of 2,202 SNPs were common to the isolates and used to classify the Nonthaburi strains as

75   members of sublineage 1.2.1. Compared with 1,601 genomes from the seven lineages of *M.*

76   *tuberculosis* complex, mutation G2342203C was found novel to the isolates in this study. Three

77   mutations (T28910C, C1180580T and C152178T) were found only in Thai Nonthaburi isolates,

78   including isolates from previous study. Although drug susceptibility tests indicated pan-

79   susceptibility, non-synonymous SNPs previously reported to be associated with resistance to

80   anti-tuberculous drugs; isoniazid, ethambutol, and ethionamide were identified in all the isolates.

81   Non-synonymous SNPs were found in virulence genes such as the genes playing roles in

82   apoptosis inhibition and phagosome arrest. We also report polymorphisms in essential genes,

83   efflux pumps associated genes and genes with known epitopes .

84   **Discussions.** The analysis of the tuberculous meningitis isolates and the availability of the

85   variations obtained will provide additional resources for global comparison of isolates from

86   pulmonary tuberculosis and tuberculosis meningitis. It will also contribute to the richness of

87   genomic databases towards the prediction of antibiotic resistance, level of virulence and of origin

88   of infection.

89

90

91

92

93

94

95   Introduction

96   Tuberculosis (TB) remains a global threat despite efforts targeted towards its control. With

97   recent advances in next generation sequencing, the analysis of bacterial whole genome sequences

98   has contributed significantly to the understanding of virulence factors and antibiotic resistance of

99   pathogenic bacteria (Koser et al. 2013; Leopold et al. 2014). Currently, there are software tools

100 and databases that are used for predicting bacterial genotype, lineages and drug resistance profile

101 from mycobacterial whole genome sequence data (Benavente et al. 2015; Coll et al. 2015).

102 Availability of more whole genome data (processed and unprocessed), especially from genotypes

103 not currently available, will contribute immensely to the profiling of pathogens.

104 Although tuberculosis is a curable disease, 9.0 million new cases and 1.5 million TB deaths were

105 recorded in 2013 (Zumla et al. 2015). This is due in part to incomplete understanding of the

106 variations that contribute to the pathogenesis and antibiotic resistance of *Mycobacterium*

107 *tuberculosis*. There are two broad types of clinical TB disease; pulmonary (PTB) in which the

108 site of infection is the lung and extra-pulmonary, including the more severe tuberculous

109 meningitis (TBM), in which the bacteria cross the blood brain barrier to get into the

110 cerebrospinal fluid (CSF) of the patient. The morbidity and mortality rate of TBM is higher than

111 PTB (Thwaites et al. 2013). The genotype of the infecting mycobacterium has been shown to be

112 one of the factors that contribute to the severity of the disease and can play a role in emergence

113 of drug resistance, susceptibility to TBM, host response and in transmissibility (Ford et al. 2013;

114 Lopez et al. 2003; Nahid et al. 2010; Thwaites et al. 2008). However the genetic factors that

115   determine the association of different lineages of mycobacteria with different level of disease

116   severity remain largely unknown.

117   There have been controversies in associating specific genotypes with morbidity or mortality from

118   TB. A study in Thailand associated the modern Beijing genotype with a more severe disease

119   progression when compared with other lineages (Faksri et al. 2011). However, in a study

120   conducted in HIV patients in Vietnam, modern Beijing genotype had lower mortality rates than

121   those infected with other lineages (Tho et al. 2012). Comparing strains isolated from TBM across

122   genotypes on a whole genome scale may provide better understanding of factors that contribute

123   to the severity of the disease.

124   IS*6110* based restriction fragment length polymorphism (RFLP) is an internationally recognized

125   method for genotyping mycobacteria (Thierry et al. 1990; van Embden et al. 1993). Nonthaburi

126   strains of *M. tuberculosis* were first identified in Thailand by its IS*6110*-RFLP patterns, usually

127   containing 9-14 bands. Subsequent spoligotyping revealed that the Nonthaburi type has a

128   spoligotype octal code 674000003413771 specifying the East-Asian India 2 Nonthaburi (EAI2-

129   Nonthaburi) genotype (Palittapongarnpim et al. 1997) [14]. It has been reported in lower

130   percentages from many countries such as the Netherlands, Australia, USA, Sweden, Saudi

131   Arabia, Tunisia, and Taiwan. However, the origin of the isolates is likely to be South East Asia,

132   as more isolates are from countries such as Indonesia, Laos PDR, Vietnam, Cambodia,

133   Philippines and Thailand (Demay et al. 2012).

134   Up to date, only relatively little information is available on the genetic characteristics of the

135   Nonthaburi strains. Three Nonthaburi strains were isolated from the CSF samples of TBM

136   patients at Siriraj Hospital, Mahidol University, Thailand. For a deeper understanding of the

137   characteristics of these isolates, genome-wide scale analysis and drug susceptibility pattern to

138   anti-tuberculosis drugs were performed and compared to the reference strain *M. tuberculosis*

139   H37Rv (NC_000962.3). The single nucleotide polymorphism (SNPs) common to the isolates

140   were compared with SNPs from 1,601 genomes from the 7 different lineages and various

141   sublineages of *M. tuberculosis* complex (MTBC). The whole genome sequence of the isolates

142   were assembled into draft genomes, annotated and have been deposited into NCBI database for

143   public access. Prior to our study, there was no complete or draft genome belonging to the

144   Nonthaburi genotype of *M. tuberculosis* in the database.

145

146   Methods

147   **Selection of strains**

148   Three isolates, CSF3053, 46-5069 and 43-13838, identified to belong to Nonthaburi genotype by

149   IS*6110*-RFLP, were selected from the stock of samples collected from the CSF of TBM patients

150   at the Drug Resistant Tuberculosis Research Fund laboratory, Department of Microbiology,

151   Faculty of Medicine Siriraj Hospital, Mahidol University, Thailand.

152   **Genomic DNA extraction**

153   Stock culture of selected strains, stored at -70°C in MH79 broth containing 15% glycerol, were

154   subcultured on Loewenstein-Jensen medium and incubated for 4 weeks at 37⁰C. DNA extraction

155   was carried out using cetyltrimethylammonium bromide (CTAB)-lysozyme enzymatic method

156   as earlier described (Larsen et al. 2007).

157   **Spoligotyping**

158   Spacer oligonucleotide typing, a polymerase chain reaction (PCR) based method used in typing

159   *M. tuberculosis* was performed following the methods earlier described (Gori et al. 2005).

160

161

162 **Whole genome sequencing and analysis**

163 Genomic DNA samples isolated from the three isolates were sequenced at Macrogen Inc., Seoul,

164 South Korea on the HiSeq 2000 platform with insert size of 300 bp (Illumina, San Diego, CA,

165 USA) yielding 100 bp paired end reads. The qualities of the sequences were assessed with

166 FastQC software (www.bioinformatics.babraham.ac.uk/projects/fastqc) to determine the

167 parameters used for trimming. Bases with quality of less than 5, reads with average of quality

168 less than 20 for every four bases, and reads with lengths that are less than 45 bases were

169 discarded using Trimmomatic software (Bolger et al. 2014) (version 0.33). The trimmed

170 sequences were aligned to the reference strain *M. tuberculosis* H37Rv (NC_000962.3) using the

171 short reads aligner, Bowtie2 (version 2.2.0) (Langmead et al. 2012). The genomic coverage was

172 estimated using Bedtools (version 2.18) (Quinlan et al. 2010). The fold coverage is estimated as

173 the number of reads supporting a particular nucleotide position on the genome. Variant calling

174 was performed on the aligned sequences using the Genome Analysis Tool Kit (GATK) (version

175 3.3) haplotype caller (McKenna et al. 2010) with minimum calling confidence threshold set at

176 phred score 30. Point allelic variation at any position within the genome when compared with the

177 reference H37Rv genome (NC_000962.3) is considered a single nucleotide polymorphism

178 (SNP).

179 SnpEFF (Cingolani et al. 2012) (version 4.0) software was used to annotate the SNPs. The

180 SNPs were filtered using standard hard filtering parameters according to GATK Best Practices

181 Recommendations (DePristo et al. 2011, Van der Auwera et al. 2013). Variants with

182 QualByDepth <2.0, FisherStrand >60, RMSMapping quality < 40,

183 MappingQualityRankSumTest < -12.5 and ReadPosRankSumTest< -8 were filtered. All SNPs

184   were confirmed using Integrated Genomic Viewer (IGV) (James et al. 2011) (version 2.0). The

185   SNPs were further grouped according to the functions of the genes in which they were found in

186   the genome when compared to the reference genome H37Rv (NC_000962.3). We evaluated

187   SNPs in groups of genes considered to be essential, drug resistance related, virulence related,

188   contain known epitopes and associated with efflux pumps.

189   The Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the

190   accession numbers LGCH00000000, LGCG00000000 and LGCF00000000. The versions

191   described in this paper are LGCH01000000, LGCG01000000 and LGCF01000000 for CSF3053,

192   46-5069 and 43-13838 respectively. The raw sequences have been deposited to the short read

193   archive (SRA) of NCBI under accession numbers SRX1094547, SRX1094546 and SRX1094545

194   for isolates CSF3053, 46-5069 and 43-13838 respectively.

195

196   **Determination of principle genetic group, lineage and sequence type**

197   Nucleotide alleles at positions 7585 and 2154724 were investigated to determine the principal

198   genetic group of the isolates as earlier defined (Sreevatsan et al. 1997). To determine the lineage

199   of the isolates, SNPs specific to different lineages as earlier reported (Coll et al. 2014b) were

200   investigated.

201   **Draft genome assembly:**

202   The paired-end raw reads of the isolates were assembled into draft genomes by using the *de novo*

203   assembly algorithm of CLC Genomics Workbench (version 7.5) which works by using a de

204   Bruijn graph (http://www.clcbio.com). The minimum contig output was set at 200 bp long.

205   Annotation of the draft genome was performed by Rapid Annotation using Subsystem

206   Technology (RAST) (http://www.nmpdr.org/) and by NCBI Prokaryotic Genome Annotation

207   Pipeline (PGAP) (http://www.ncbi.nlm.nih.gov/genome/annotation_prok/).

208

209   **Comparison of Nonthaburi isolates with isolates from other lineages**

210   The SNPs that are common to the three isolates were compared with 92,000 SNPs from 1,601

211   genomes of MTBC previously reported (Coll et al. 2014a)

212   (http://pathogenseq.lshtm.ac.uk/phytblive/index.php). These include 121, 390, 189, 856, 17, 11,

213   and 6 genomes from lineages 1, 2, 3, 4, 5, 6, and 7 respectively. Eleven samples from *M. bovis*

214   were also included.

215   **Large sequence polymorphism determination**

216   Regions of differences when compared with reference strain H37Rv (NC_000926.3) were

217   determined by using the indel and structural variants determination tool of CLC Genomics

218   Workbench (version 7.5) (http://www.clcbio.com) and Bedtools (version 2.18) (Quinlan et al.

219   2010). The regions of deletions were confirmed with PCR using primers CF

220   (CATCCGCACCGAACCTGTAA) and CR (AACCGTTCACGACAAGCAAC), AF

221   (GCCCAACCTGATTGGTTTCG) and AR (CAAACGCTCGCCATGATCTC), BF

222   (TCGACTGCCATACAACCTGC) and BR (ACTTCCGGTGGTAACAGTGC) respectively for

223   RD239, RD147c and newly identified deletion of 500 bp between 3501224-3501724 (*M.*

224   *tuberculosis* H37Rv (NC_000962.3 genome numbering). The reactions were performed with

225   initial denaturation at 94 °C and 30 cycles of denaturation for 1 minute, annealing of primers at

226   60 °C for 1 minute and extension with platinum *Taq* DNA polymerase for 1 minute at 68 °C.

227   Final extensions were performed at 68 °C for 10 minutes. The reactions were performed as

228   recommended by the manufacturer of the DNA polymerase.

229

230 **Drug susceptibility testing**

231 The susceptibility of the isolates to first line drugs and other second-line anti-tuberculosis drugs

232 was investigated using the standard agar proportion method (Larsen et al. 2007). The drug

233 concentrations used in the test comprise 0.2 mg/l isoniazid, 1.0 mg/l rifampicin, 2.0 mg/l

234 streptomycin, 5.0 mg/l ethambutol, 1.0 mg/l linezolid, 6.0 mg/l amikacin, 5.0 mg/l ethionamide,

235 2.0 mg/l paraaminosalycic acid, 2.0 mg/ml ofloxacin, 2.0 mg/l moxifloxacin, 2.0 mg/l

236 gatifloxacin, 1.0 mg/ml sitafloxacin, 6.0 mg/l kanamycin, 2.0 mg/l ciprofloxacin, 2.0 mg/l

237 levofloxacin, and 3.0 mg/l clarithromycin. Growth equal to or more than 1% on drug containing

238 media compared to drug free media was recorded as drug resistance. The phenotypic drug testing

239 was performed on the initial isolates from the patients and repeated on the stock cultures.

240 **Ethical Approval**

241 The study was approved by the Institutional review board (IRB) of Faculty of Medicine Siriraj

242 Hospital, Mahidol University SiEC No. 152/2549

243

244 Results and discussion

245 For the three isolates CSF3053, 46-5069 and 43-13838, an average of 99.1% of raw reads

246 mapped to the reference genome. On the average, 99.8% of the reference was covered to at least

247 1-fold coverage. The depth across all the positions covered by the reads was about 1,056-fold on

248 the average, Table 1.

249

250

251

252   **Genome assembly**

253   The sequences of the isolates were assembled and annotated as described in Methods. 159

254   contigs with N_{50} of 69,028, 173 contigs with N_{50} of 63,852, and 177 contigs with N_{50} of 63,019

255   contigs were obtained for CSF3053, 43-5069 and 46-13838 respectively. All isolates have 65.5

256   % guanine/cytosine (GC) content, typical of mycobacteria. The draft genomes have an average

257   size of 4,364,461 bp, 4,154 genes, 48 rRNAs and 64 pseudogenes. Details of the assembly and

258   annotation are shown in Table 1.

259

260   **Single nucleotide polymorphisms**

261   Point allelic variations at any position within the genome when compared with the reference

262   H37Rv genome (NC_000962.3) were investigated.

263   In total, 2,202 positions were found to have similar allelic changes (SNPs) in all isolates as

264   shown in Figure 1. 1,963 are in coding regions (754 synonymous, 1209 (61.6%) non

265   synonymous) and 239 are intergenic. In this study, CSF3053, 46-5069 and 43-13838 have 10, 7

266   and 49 unique SNPs respectively. 43-13838 and CSF3053 have 23 SNPs in common, CSF3053

267   and 46-5069 have 99 SNPs in common, while 43-13838 and 46-5069 have 7 SNPs in common.

268   Using the SNPs, the isolates were found to belong to lineage 1 with the presence of allele C/A

269   and G/C at positions 2154724 and 7585 resulting in *katG* R463L and *gyrA* S95T respectively

270   (Sreevatsan et al. 1997). Using a recently developed SNP barcode (Coll et al. 2014a), the isolates

271   were found to be specific to Indo Oceanic lineage 1.2.1, with nucleotide changes G/A at position

272   615938, C/A at position 3479545, G/C at position 4244420 and G/C at position 9260.

273   The 2,202 SNPs that were found to be common to the isolates in this study were compared with

274   92,000 SNPs from 1,601 genomes of MTBC that were analyzed previously. These include 121,

275 390, 189, 856, 17, 11, and 6 genomes from lineages 1, 2, 3, 4, 5, 6, and 7, respectively. Eleven

276 samples from *M. bovis* were also included (Coll et al. 2014a). The common SNPs were used to

277 position the strains on a phylogenetic tree compared to other strains and lineages of MTBC as

278 shown in figure S1. Nucleotide change G/C at position 2342203 was found only in the isolates in

279 this study when compared with the 1,601 MTBC genomes. There is evidence from macrophage

280 systems that strain-to-strain variability affects phenotypic outcomes (McEvoy et al. 2012).

281 Phylogeographic strain variation may therefore have considerable effect on the development of

282 new diagnostic tools, vaccines and drugs.

283 SNP C/T at position 3378828 was reported to be unique to members of lineage 1 (Coll et al.

284 2014a). Although this SNP was found in many genomes belonging to lineage 1, we found out

285 that it was absent in the three isolates in this study and in 6 other Nonthaburi isolates from

286 Thailand and the Netherlands used in previous studies which are grouped under lineage 1. This

287 indicates that the allele change at this position may be specific only to a sub-branch of lineage 1.

288 Synonymous SNP T/C at position 28910, non-synonymous SNP C/T at position 152178

289 resulting in Thr344Ile in *pepA* gene and intergenic SNP C/T at position 1180580 were found

290 only in Nonthaburi isolates from Thailand. They were not found in any genome belonging to

291 lineages 2, 3, 4, 5, 6 and 7. Within lineage 1, these SNPs were found only in Thai Nonthaburi

292 isolates, from previous study (Coll et al. 2014a)*,* and the isolates in this study. They were

293 however absent in the Nonthaburi genotype isolates from the Netherlands. *pepA* gene is a

294 probable serine protease with the exact function unknown. It is in the intermediary metabolism

295 and respiration functional category. Its mRNA was found to be upregulated after 96 hours of

296 starvation (Betts et al. 2002), suggesting its role in the adaptation of mycobacteria to extreme

297   conditions. The association of the SNPs at these positions with Thailand warrants further

298   investigation.

299

300

301   **Large sequence polymorphism**

302   Region of difference RD239 that is specific to lineage 1 of MTBC and previously reported

303   RD147c, not specific to lineage 1, were found in all the three isolates. In addition, a region of

304   deletion of 500 bp between 3501224-3501724 (*M. tuberculosis* H37Rv (NC_000962.3 genome

305   numbering) comprising Rv3135 (*ppe50*), was observed in all isolates. The details of the deletions

306   as well as the affected open reading frames are shown in Table 2. The deletions were confirmed

307   with PCR (see Figures S2, S3 and S4). The PE-PPE protein class, while not well characterized,

308   represents the third most abundant category of mycobacterial proteins and showed the most

309   consistent expression during infection (Kruh et al. 2010). Although PPE50 has a yet unknown

310   function, it was listed among promising therapeutic target in tuberculosis treatment based on its

311   expression, and homology to human and other microbial proteins (Raman et al. 2008). The

312   deletion of this gene may be a means of evading recognition by the host immune system.

313   Deletions have been shown to have a wide range of effects on *M. tuberculosis* including

314   association with an increased probability of transmission (Tsolaki et al. 2004).

315

316   **Polymorphisms in drug resistance associated genes**

317   Despite being isolated from patients with severe form of tuberculosis, drug susceptibility tests

318   results show that the three isolates are susceptible to first line drugs; isoniazid, rifampicin,

319 ethambutol and streptomycin, and to quinolones: ciprofloxacin, ofloxacin, gatifloxacin,

320 moxifloxacin, levofloxacin, and sitafloxacin. They were also found to be susceptible to linezolid,

321 amikacin, ethionamide, paraaminosalicylic acid, kanamycin and clarithromycin.

322 However, 37 SNPs were found in drug resistant related genes reported in TBdream database and

323 other earlier published reports (Sandgren et al. 2009). Nineteen are synonymous while 18 are non

324 synonymous. Non synonymous mutations Gly312Ser of *kasA* gene and Ile73Thr in *efpA* were

325 previously reported to be associated with isoniazid resistance (Mdluli et al. 1998; Ramaswamy et

326 al. 2003), but were found in our isolates. Association between these mutations and resistance to

327 isoniazid needs to be confirmed. *iniA* gene and Rv1592c were reported to be associated with

328 tolerance to isoniazid (Colangeli et al. 2005; Ramaswamy et al. 2003). In our analysis, mutations

329 His481Gln in *iniA* gene and Ile322Val in *Rv1592c* were found. These positions may not be

330 associated with the supposed roles of these genes in isoniazid resistance.

331 Polymorphism exists at position 237 of *nudC* in *M. tuberculosis* isolates (Wang et al. 2011). In

332 particular, the amino acid change Gln237Pro in *nudC* is found in the Indo Oceanic and West

333 African lineages. It was demonstrated to prevent dimer formation and results in the loss of

334 activity of the enzyme. It was also shown to degrade the active forms of isoniazid and

335 ethionamide (Wang et al. 2011). We however found this codon change in all isolates in this

336 study. This suggests the non-involvement of the amino acid change at this position in resistance

337 to both drugs.

338 Mutations Cys110Tyr in *embR*, Thr270Ile and Asn394Asp in *embC*, Pro913Ser in *embA* and

339 Glu378Ala in *embB*, were previously reported to be involved in ethambutol resistance

340 (Ramaswamy et al. 2000; Srivastava et al. 2009). However, these mutations were found in this

341 study. Mutation Ser257Pro in *rmlD* was suspected to be involved in isoniazid and ethambutol

342   resistance (Ramaswamy et al. 2000). This was however found in all isolates considered in this

343   study. Mutations Glu21Gln in *gyrA*, Ile322Val in Rv1592c, Arg463Leu in *katG*, and Arg93Leu

344   in *cycA* were found to be common to the isolates in this study. They have also been reported to

345   be common to pan-susceptible and drug resistant *M. tuberculosis* sequence type 10 Beijing

346   isolates (Regmi et al. 2015). Our results confirm that these mutations are polymorphic rather than

347   being involved in drug resistance. The details of the synonymous and non-synonymous SNPs

348   found in drug resistant related genes and the predicted protein variation effects are shown in

349   Table 3.

350

351   **Polymorphisms in virulence genes, efflux pump related genes, and essential genes**

352   Oftentimes, mutations provide selective advantage to an organism in a particular environment.

353   Some non-synonymous mutations in *rpoC* gene have been shown to result in higher

354   competitiveness *in vitro* and have higher fitness *in vivo* evidenced by their prevalence across

355   patient populations (Comas et al. 2012). In this study, we found Ala172Val mutation in *rpoC*

356   gene in all isolates.

357   We also sought to determine polymorphisms in genes that play important roles in the survival

358   and pathogenesis of *M. tuberculosis.* Of particular interest are the genes that are involved in the

359   evasion of the host immune system. SNPs in 37 mycobacteria virulence related genes were found

360   to be common to the isolates. Twenty nine of the SNPs are non-synonymous. Polyketide

361   synthases (PKs) are group of genes involved in the synthesis of polyketides which are

362   structurally complex compounds produced by organisms for survival advantage. Some

363   mycobacteria PKs genes such as *pks15, pks1, pks10, pks12, pks5,* and *pks7* are known to be

364   involved in virulence (Reed et al. 2004; Rousseau et al. 2003; Sirakova et al. 2003; Tsenova et

365 al. 2005). Insertion of 7 base pairs was found in *pks15/1* junction in all isolates. The presence of

366 the 7 base pair insertion leads to a frame shift that results in the loss of stop codon of *pks15*. This

367 results in a continuous transcription of *pks15* and *pks1*. This was previously associated with the

368 more virulent phenotype of the modern Beijing family, but such claim has since been refuted as

369 it can be found across the seven lineages. The implication of the insertion needs further

370 experiments to understand. Two mutations Ile474Met and Thr604Ala were found in *nuoG* gene.

371 *nuoG* is a probable NADH dehydrogenase, reported to be involved in apoptosis inhibition

372 (Velmurugan et al. 2007). Mutation Arg463Leu was found in *katG*, a gene previously implicated

373 in inhibiting antimicrobial effectors of the macrophage (Ng et al. 2004). Protein kinases such as

374 *pknD* and *pknG* are important virulent factors of *M. tuberculosis*. *pknD* has been reported to play

375 a role in the infection of the host's central nervous system by *M. tuberculosis* (Be et al. 2012;

376 Cowley et al. 2004). Gln472Pro mutation in *pknD* was found in all isolates. *virS* is a

377 transcription regulator that belongs to AraC family. Its attenuation in a mouse model resulted in

378 an increased animal survival (Gupta et al. 1999; Singh et al. 2003). We found mutation

379 Leu316Arg in this gene in all isolates.

380 Stop codon was gained after Arg305 in *PStA1,* an inorganic phosphate ABC transporter. Stop

381 codon was however lost in *Rv1504*. The stop codon was replaced with glutamine as codon 200.

382 *Rv1504* and *PstA1* were reported to be involved in the adaptation and survival of mycobacteria in

383 macrophages (Brodin et al. 2010; Rengarajan et al. 2005).

384 Non-synonymous and synonymous SNPs were found in other genes involved in various other

385 functions related to virulence such as synthesis of complex and simple lipids, cell wall proteins,

386 lipoproteins, cholesterol metabolism, secretion systems, protein kinases, metal transporter

387 proteins, two component systems and other proteins of unknown functions (Table S1).

388 Efflux pumps play roles in drug resistance, cell physiology, detoxification and virulence of *M.*

389 *tuberculosis* (Nikaido 2009). Ten synonymous SNPs and 15 non-synonymous SNPs were found

390 in efflux pump related genes. One stop codon was gained by *Rv2994*, a predicted transmembrane

391 protein involved in efflux system (Table S2).

392

393 Twenty eight SNPs were observed in genes with known epitopes, 11 are synonymous while 17

394 are non-synonymous (Table S3).

395 In addition, 316 SNPs were found in essential genes, 135 are synonymous, 181 are non-

396 synonymous. A start codon was lost in *pabB* gene. *pabB* is a cell membrane associated gene that

397 encodes *para*-aminobenzoate synthetase component-I involved in the biosynthesis of *p*-

398 aminobenzoate, a precursor of folate biosynthesis (Sassetti et al. 2003; Zheng et al. 2008). The

399 details of the position, nucleotide change, amino acid change and the genes involved are

400 presented in Table S4.

401 The association of the SNPs or deletions reported in this study to TBM needs further

402 investigations. This can be done by comparing them with variations from PTB cases, to

403 determine exclusive associations with TBM. Furthermore, The involvement of the reported

404 allelic changes in the functions of the various genes from which they were found can be verified

405 by site directed mutagenesis in laboratory strains of *M. tuberculosis*, and subsequent animal

406 experiments.

407 Conclusion

408 Genetic factors that contribute to the ability of infecting mycobacteria in causing TBM remain

409 largely unknown. We have presented a detailed analysis of the polymorphism existing in the

410 genome of Nonthaburi isolates from TBM patients, when compared to reference strain *M.*

411   *tuberculosis* H37Rv (NC_000962.3). The polymorphisms were compared to 1,601 genomes

412   representing the members of the 7 MTBC lineages. Uniqueness of certain SNPs to certain

413   genotypes, countries or region such as found in this study may be useful epidemiologically to

414   determine the origin of an infection and potential level of disease severity. We have also

415   presented the first draft genomes of *M. tuberculosis* Nonthaburi genotype.

416   Many studies have reported the SNPs playing roles in drug resistance in many drug resistant

417   related genes. These have majorly formed the basis for the development of some databases. It is

418   equally important to report polymorphisms found in these genes from drug susceptible strains so

419   that SNPs that are not involved in resistance to drugs but present in the drug resistance related

420   genes could be filtered out in the process of predicting drug resistance. Our results will also form

421   a basis for comparison with other genotypes of mycobacteria isolated from the CSF of TBM or

422   sputum of PTB patients in order to identify potential factors contributing to TBM.

423

424   **Supplementary Information:**

425   Supplementary Figures S2, S3, S4

426   Supplementary Tables S1, S2

427   Supplementary Table S3

428   Supplementary Table S4

429

430

431

432   References

433 Be NA, Bishai WR, and Jain SK. 2012. Role of *Mycobacterium tuberculosis* pknD in the
434   pathogenesis of central nervous system tuberculosis. *BMC Microbiology* 12:7.
435   10.1186/1471-2180-12-7
436 Benavente ED, Coll F, Furnham N, McNerney R, Glynn JR, Campino S, Pain A, Mohareb FR,
437   and Clark TG. 2015. PhyTB: Phylogenetic tree visualisation and sample positioning
438   for *M. tuberculosis*. *BMC Bioinformatics* 16:155. 10.1186/s12859-015-0603-3
439 Betts JC, Lukey PT, Robb LC, McAdam RA, and Duncan K. 2002. Evaluation of a nutrient
440   starvation model of *Mycobacterium tuberculosis* persistence by gene and protein
441   expression profiling. *Molecular Microbiology* 43:717-731.
442 Bolger AM, Lohse M and Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina
443 Sequence Data. *Bioinformatics*, btu170
444
445 Brodin P, Poquet Y, Levillain F, Peguillet I, Larrouy-Maumus G, Gilleron M, Ewann F,
446   Christophe T, Fenistein D, Jang J, Jang MS, Park SJ, Rauzier J, Carralot JP, Shrimpton
447   R, Genovesio A, Gonzalo-Asensio JA, Puzo G, Martin C, Brosch R, Stewart GR, Gicquel
448   B, and Neyrolles O. 2010. High content phenotypic cell-based visual screen identifies
449   *Mycobacterium tuberculosis* acyltrehalose-containing glycolipids involved in
450   phagosome remodeling. *PLoS Pathogens* 6:e1001100.
451   10.1371/journal.ppat.1001100
452 1. Choi Y, Sims GE, Murphy S, Miller JR and Chan AP. 2012. Predicting the Functional Effect of
453 Amino Acid Substitutions and Indels. *PLoS ONE* 7(10): e46688
454
455 Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X and
456   Ruden DM. 2012. A program for annotating and predicting the effects of
457   single nucleotide polymorphisms, SnpEff: SNPs in the genome of
458   Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly 6(2):80-92.*
459
460 Colangeli R, Helb D, Sridharan S, Sun J, Varma-Basil M, Hazbon MH, Harbacheuski R,
461   Megjugorac NJ, Jacobs WR, Jr., Holzenburg A, Sacchettini JC, and Alland D. 2005. The
462   *Mycobacterium tuberculosis iniA* gene is essential for activity of an efflux pump that
463   confers drug tolerance to both isoniazid and ethambutol. *Molecular Microbiology*
464   55:1829-1840. 10.1111/j.1365-2958.2005.04510.x
465 Coll F, McNerney R, Guerra-Assuncao JA, Glynn JR, Perdigao J, Viveiros M, Portugal I, Pain A,
466   Martin N, and Clark TG. 2014a. A robust SNP barcode for typing *Mycobacterium*
467   *tuberculosis* complex strains. *Nature Communications* 5:4812.
468   10.1038/ncomms5812
469 Coll F, McNerney R, Preston MD, Guerra-Assuncao JA, Warry A, Hill-Cawthorne G, Mallard
470   K, Nair M, Miranda A, Alves A, Perdigao J, Viveiros M, Portugal I, Hasan Z, Hasan R,
471   Glynn JR, Martin N, Pain A, and Clark TG. 2015. Rapid determination of anti-
472   tuberculosis drug resistance from whole-genome sequences. *Genome Medicine* 7:51.
473   10.1186/s13073-015-0164-0
474 Coll F, Preston M, Guerra-Assuncao JA, Hill-Cawthorn G, Harris D, Perdigao J, Viveiros M,
475   Portugal I, Drobniewski F, Gagneux S, Glynn JR, Pain A, Parkhill J, McNerney R,
476   Martin N, and Clark TG. 2014b. PolyTB: a genomic variation map for *Mycobacterium*
477   *tuberculosis*. *Tuberculosis* 94:346-354. 10.1016/j.tube.2014.02.005

478 Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, Galagan J, Niemann S, and
479   Gagneux S. 2012. Whole-genome sequencing of rifampicin-resistant *Mycobacterium*
480   *tuberculosis* strains identifies compensatory mutations in RNA polymerase genes.
481   *Nature Genetics* 44:106-110. 10.1038/ng.1038.
482 Cowley S, Ko M, Pick N, Chow R, Downing KJ, Gordhan BG, Betts JC, Mizrahi V, Smith DA,
483   Stokes RW, and Av-Gay Y. 2004. The *Mycobacterium tuberculosis* protein
484   serine/threonine kinase PknG is linked to cellular glutamate/glutamine levels and is
485   important for growth in vivo. *Molecular Microbiology* 52:1691-1702.
486   10.1111/j.1365-2958.2004.04085.x
487
488 DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G,
489 Rivas MA, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K, Gabriel
490 S, Altshuler D and Daly M. 2011. A framework for variation discovery and genotyping using
491 next-generation DNA sequencing data. *Nature Genetics* 43:491-498
492
493 Demay C, Liens B, Burguiere T, Hill V, Couvin D, Millet J, Mokrousov I, Sola C, Zozio T, and
494   Rastogi N. 2012. SITVITWEB--a publicly available international multimarker
495   database for studying *Mycobacterium tuberculosis* genetic diversity and molecular
496   epidemiology. *Infection, genetics and evolution : Journal of Molecular Epidemiology*
497   *and Evolutionary Genetics in Infectious Diseases* 12:755-766.
498   10.1016/j.meegid.2012.02.004
499 Faksri K, Drobniewski F, Nikolayevskyy V, Brown T, Prammananan T, Palittapongarnpim P,
500   Prayoonwiwat N, and Chaiprasert A. 2011. Epidemiological trends and clinical
501   comparisons of *Mycobacterium tuberculosis* lineages in Thai TB meningitis.
502   *Tuberculosis* 91:594-600. 10.1016/j.tube.2011.08.005
503 Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, Johnston JC, Gardy J,
504   Lipsitch M, and Fortune SM. 2013. *Mycobacterium tuberculosis* mutation rate
505   estimates from different lineages predict substantial differences in the emergence of
506   drug-resistant tuberculosis. *Nature Genetics* 45:784-790. 10.1038/ng.2656
507 Gori A, Bandera A, Marchetti G, Degli Esposti A, Catozzi L, Nardi GP, Gazzola L, Ferrario G,
508   van Embden JD, van Soolingen D, Moroni M, and Franzetti F. 2005. Spoligotyping
509   and *Mycobacterium tuberculosis*. *Emerging Infectious Diseases* 11:1242-1248.
510   10.3201/eid1108.040982.
511 Guillemin I, Jarlier V, and Cambau E. 1998. Correlation between Quinolone Susceptibility
512   Patterns and Sequences in the A and B Subunits of DNA Gyrase in Mycobacteria.
513   *Antimicrobial Agents and Chemotherapy* 42:2084-2088
514
515 Gupta S, Jain S, and Tyagi AK. 1999. Analysis, expression and prevalence of the
516   *Mycobacterium tuberculosis* homolog of bacterial virulence regulating proteins.
517   *FEMS Microbiology Letters* 172:137-143
518 Heym B, Alzari PM, Honore N, and Cole ST. 1995. Missense mutations in the catalase-
519   peroxidase gene, katG, are associated with isoniazid resistance in *Mycobacterium*
520   *tuberculosis*. *Molecular Microbiology* 15:235-245
521
522 James TR, Helga T, Wendy W, Mitchell G, Eric SL ,Gad G, and Jill PM. 2011.

523   Integrative Genomics Viewer. *Nature Biotechnology*. 29, 24-26.

524 Kapur V, Li LL, Hamrick MR, Plikaytis BB, Shinnick TM, Telenti A, Jacobs WR, Jr., Banerjee A,
525   Cole S and Yuen KY. 1995. Rapid Mycobacterium species assignment and
526   unambiguous identification of mutations associated with antimicrobial resistance in
527   *Mycobacterium tuberculosis* by automated DNA sequencing. *Archives of Pathology &*
528   *Laboratory Medicine* 119:131-138

529

530 Koser CU, Bryant JM, Becq J, Torok ME, Ellington MJ, Marti-Renom MA, Carmichael AJ,
531   Parkhill J, Smith GP, and Peacock SJ. 2013. Whole-genome sequencing for rapid
532   susceptibility testing of *M. tuberculosis. The New England Journal of Medicine*
533   369:290-292. 10.1056/NEJMc1215305.

534 Kruh NA, Troudt J, Izzo A, Prenni J, and Dobos KM. 2010. Portrait of a pathogen: the
535   *Mycobacterium tuberculosis* proteome in vivo. *PloS One* 5:e13938.
536   10.1371/journal.pone.0013938

537 Langmead B and Salzberg S. 2012. Fast gapped-read alignment with Bowtie 2. *Nature*
538 *Methods*. 9:357-359

539

540 Larsen MH, Biermann K, Tandberg S, Hsu T, and Jacobs WR, Jr. 2007. Genetic Manipulation
541   of *Mycobacterium tuberculosis. Curr Protoc Microbiol* Chapter 10:Unit 10A 12.
542   10.1002/9780471729259.mc10a02s6

543 Lavender C, Globan M, Sievers A, Billman-Jacobe H, and Fyfe J. 2005. Molecular
544   characterization of isoniazid-resistant *Mycobacterium tuberculosis* isolates collected
545   in Australia. *Antimicrobial Agents and Chemotherapy* 49:4068-4074.
546   10.1128/AAC.49.10.4068-4074.2005

547 Lee AS, Lim IH, Tang LL, Telenti A, and Wong SY. 1999. Contribution of kasA analysis to
548   detection of isoniazid-resistant *Mycobacterium tuberculosis* in Singapore.
549   *Antimicrobial Agents and Chemotherapy* 43:2087-2089

550

551 Leopold SR, Goering RV, Witten A, Harmsen D, and Mellmann A. 2014. Bacterial whole-
552   genome sequencing revisited: portable, scalable, and standardized analysis for
553   typing and detection of virulence and antibiotic resistance genes. *Journal of Clinical*
554   *Microbiology* 52:2365-2370. 10.1128/JCM.00262-14

555 Lopez B, Aguilar D, Orozco H, Burger M, Espitia C, Ritacco V, Barrera L, Kremer K,
556   Hernandez-Pando R, Huygen K, and van Soolingen D. 2003. A marked difference in
557   pathogenesis and immune response induced by different *Mycobacterium*
558   *tuberculosis* genotypes. *Clinical and Experimental Immunology* 133:30-37

559 Mathys V, Wintjens R, Lefevre P, Bertout J, Singhal A, Kiass M, Kurepina N, Wang XM,
560   Mathema B, Baulard A, Kreiswirth BN, and Bifani P. 2009. Molecular genetics of
561   para-aminosalicylic acid resistance in clinical isolates and spontaneous mutants of
562   *Mycobacterium tuberculosis. Antimicrobial Agents and Chemotherapy* 53:2100-2109.
563   10.1128/AAC.01197-08

564

565 McEvoy CR, Cloete R, Muller B, Schurch AC, van Helden PD, Gagneux S, Warren RM, and Gey
566   van Pittius NC. 2012. Comparative analysis of *Mycobacterium tuberculosis* pe and
567   ppe genes reveals high sequence variation and an apparent absence of selective
568   constraints. *PloS One* 7:e30593. 10.1371/journal.pone.0030593

569 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella
570   K, Altshuler D, Gabriel S, Daly M and DePristo MA. 2010. The Genome
571   Analysis Toolkit: a MapReduce framework for analyzing next-generation
572   DNA sequencing data. *Genome Research* 20:1297-303
573
574 Mdluli K, Slayden RA, Zhu Y, Ramaswamy S, Pan X, Mead D, Crane DD, Musser JM, and Barry
575   CE, 3rd. 1998. Inhibition of a *Mycobacterium tuberculosis* beta-ketoacyl ACP
576   synthase by isoniazid. *Science* 280:1607-1610.
577 Nahid P, Bliven EE, Kim EY, Mac Kenzie WR, Stout JE, Diem L, Johnson JL, Gagneux S,
578   Hopewell PC, Kato-Maeda M, and Tuberculosis Trials C. 2010. Influence of *M.*
579   *tuberculosis* lineage variability within a clinical trial for pulmonary tuberculosis.
580   *PloS One* 5:e10753. 10.1371/journal.pone.0010753
581 Ng VH, Cox JS, Sousa AO, MacMicking JD, and McKinney JD. 2004. Role of KatG catalase-
582   peroxidase in mycobacterial pathogenesis: countering the phagocyte oxidative
583   burst. *Molecular Microbiology* 52:1291-1302. 10.1111/j.1365-2958.2004.04078.x
584 Nikaido H. 2009. Multidrug resistance in bacteria. *Annual Review of Biochemistry* 78:119-
585   146. 10.1146/annurev.biochem.78.082907.145923
586 Okamoto S, Tamaru A, Nakajima C, Nishimura K, Tanaka Y, Tokuyama S, Suzuki Y, and Ochi
587   K. 2007. Loss of a conserved 7-methylguanosine modification in 16S rRNA confers
588   low-level streptomycin resistance in bacteria. *Molecular Microbiology* 63:1096-
589   1106. 10.1111/j.1365-2958.2006.05585.x
590
591 Palittapongarnpim P, Luangsook P, Tansuphaswadikul S, Chuchottaworn C, Prachaktam R,
592   and Sathapatayavongs B. 1997. Restriction fragment length polymorphism study of
593   *Mycobacterium tuberculosis* in Thailand using IS6110 as probe. *The International*
594   *Journal of Tuberculosis and Lung Disease : the Official Journal of the International*
595   *Union Against Tuberculosis and Lung Disease* 1:370-376
596 Projahn M, Koser CU, Homolka S, Summers DK, Archer JA, and Niemann S. 2011.
597   Polymorphisms in isoniazid and prothionamide resistance genes of the
598   *Mycobacterium tuberculosis* complex. *Antimicrobial Agents and Chemotherapy*
599   55:4408-4411. 10.1128/AAC.00555-11
600
601 Quinlan AR and Hall IM. 2010. BEDTools: a flexible suite of utilities for
602   comparing genomic features. *Bioinformatics.* 26, 6, pp. 841–842
603
604 Raman K, Yeturu K, and Chandra N. 2008. targetTB: a target identification pipeline for
605   *Mycobacterium tuberculosis* through an interactome, reactome and genome-scale
606   structural analysis. *BMC Systems Biology* 2:109. 10.1186/1752-0509-2-109
607 Ramaswamy SV, Amin AG, Goksel S, Stager CE, Dou SJ, El Sahly H, Moghazeh SL, Kreiswirth
608   BN, and Musser JM. 2000. Molecular genetic analysis of nucleotide polymorphisms
609   associated with ethambutol resistance in human isolates of *Mycobacterium*
610   *tuberculosis*. *Antimicrobial Agents and Chemotherapy* 44:326-336.

611   Ramaswamy SV, Reich R, Dou SJ, Jasperse L, Pan X, Wanger A, Quitugua T, and Graviss EA.
612   2003. Single nucleotide polymorphisms in genes associated with isoniazid
613   resistance in *Mycobacterium tuberculosis. Antimicrobial Agents and Chemotherapy*
614   47:1241-1250.
615   Reed MB, Domenech P, Manca C, Su H, Barczak AK, Kreiswirth BN, Kaplan G, and Barry CE,
616   3rd. 2004. A glycolipid of hypervirulent tuberculosis strains that inhibits the innate
617   immune response. *Nature* 431:84-87. 10.1038/nature02837
618   Regmi SM, Coker OO, Kulawonganunchai S, Tongsima S, Prammananan T, Viratyosin W,
619   Thaipisuttikul I, and Chaiprasert A. 2015. Polymorphisms in drug-resistant-related
620   genes shared among drug-resistant and pan-susceptible strains of sequence type 10,
621   Beijing family of *Mycobacterium tuberculosis. International Journal of*
622   *Mycobacteriology* 4:67-72. http://dx.doi.org/10.1016/j.ijmyco.2014.11.050
623   Rengarajan J, Bloom BR, and Rubin EJ. 2005. Genome-wide requirements for
624   *Mycobacterium tuberculosis a*daptation and survival in macrophages. *Proceedings of*
625   *the National Academy of Sciences of the United States of America* 102:8327-8332.
626   10.1073/pnas.0503272102
627   Rousseau C, Sirakova TD, Dubey VS, Bordat Y, Kolattukudy PE, Gicquel B, and Jackson M.
628   2003. Virulence attenuation of two Mas-like polyketide synthase mutants of
629   *Mycobacterium tuberculosis. Microbiology* 149:1837-1847
630
631   Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM and Murray MB.
632   2009. *Tuberculosis Drug Resistance Mutation Database. PLoS Medicine*
633   6(2): e1000002. doi:10.1371/journal.pmed.1000002
634
635   Sassetti CM, Boyd DH, and Rubin EJ. 2003. Genes required for mycobacterial growth
636   defined by high density mutagenesis. *Molecular Microbiology* 48:77-84
637
638   Siddiqi N, Das R, Pathak N, Banerjee S, Ahmed N, Katoch VM, and Hasnain SE. 2004.
639   *Mycobacterium tuberculosis* isolate with a distinct genomic identity overexpresses a
640   tap-like efflux pump. *Infection* 32:109-111. 10.1007/s15010-004-3097-x
641
642   Singh A, Jain S, Gupta S, Das T, and Tyagi AK. 2003. mymA operon of Mycobacterium
643   tuberculosis: its regulation and importance in the cell envelope. *FEMS Microbiology*
644   *Letters* 227:53-63.
645   Sirakova TD, Dubey VS, Kim HJ, Cynamon MH, and Kolattukudy PE. 2003. The largest open
646   reading frame (pks12) in the *Mycobacterium tuberculosis* genome is involved in
647   pathogenesis and dimycocerosyl phthiocerol synthesis. *Infection and Immunity*
648   71:3794-3801.
649   Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, Whittam TS, and Musser
650   JM. 1997. Restricted structural gene polymorphism in the *Mycobacterium*
651   *tuberculosis* complex indicates evolutionarily recent global dissemination.
652   *Proceedings of the National Academy of Sciences of the United States of America*
653   94:9869-9874.
654   Srivastava S, Ayyagari A, Dhole TN, Nyati KK, and Dwivedi SK. 2009. emb nucleotide
655   polymorphisms and the role of embB306 mutations in *Mycobacterium tuberculosis*

656   resistance to ethambutol. *International Journal of Medical Microbiology : IJMM*
657   299:269-280. 10.1016/j.ijmm.2008.07.001
658 Taniguchi H, Aramaki H, Nikaido Y, Mizuguchi Y, Nakamura M, Koga T, and Yoshida S. 1996.
659   Rifampicin resistance and mutation of the rpoB gene in *Mycobacterium tuberculosis*.
660   *FEMS Microbiology Letters* 144:103-108.
661 Telenti A, Philipp WJ, Sreevatsan S, Bernasconi C, Stockbauer KE, Wieles B, Musser JM, and
662   Jacobs WR, Jr. 1997. The emb operon, a gene cluster of *Mycobacterium tuberculosis*
663   involved in resistance to ethambutol. *Nature Medicine* 3:567-570.
664
665 Thierry D, Cave MD, Eisenach KD, Crawford JT, Bates JH, Gicquel B, and Guesdon JL. 1990.
666   IS6110, an IS-like element of *Mycobacterium tuberculosis* complex. *Nucleic acids*
667   *research* 18:188.
668 Tho DQ, Torok ME, Yen NT, Bang ND, Lan NT, Kiet VS, van Vinh Chau N, Dung NH, Day J,
669   Farrar J, Wolbers M, and Caws M. 2012. Influence of antituberculosis drug resistance
670   and *Mycobacterium tuberculosis* lineage on outcome in HIV-associated tuberculous
671   meningitis. *Antimicrobial Agents and Chemotherapy* 56:3074-3079.
672   10.1128/AAC.00319-12
673 Thwaites G, Caws M, Chau TT, D'Sa A, Lan NT, Huyen MN, Gagneux S, Anh PT, Tho DQ,
674   Torok E, Nhu NT, Duyen NT, Duy PM, Richenberg J, Simmons C, Hien TT, and Farrar
675   J. 2008. Relationship between *Mycobacterium tuberculosis* genotype and the clinical
676   phenotype of pulmonary and meningeal tuberculosis. *Journal of Clinical*
677   *Microbiology* 46:1363-1368. 10.1128/JCM.02180-07
678 Thwaites GE , van Toorn R and Schoeman J. 2013. Tuberculous meningitis: more questions,
679 still too few answers. Lancet Neurol 12:999-1010
680 Tsenova L, Ellison E, Harbacheuski R, Moreira AL, Kurepina N, Reed MB, Mathema B, Barry
681   CE, 3rd, and Kaplan G. 2005. Virulence of selected *Mycobacterium tuberculosis*
682   clinical isolates in the rabbit model of meningitis is dependent on phenolic
683   glycolipid produced by the bacilli. *The Journal of Infectious Diseases* 192:98-106.
684   10.1086/430614
685 Tsolaki AG, Hirsh AE, DeRiemer K, Enciso JA, Wong MZ, Hannan M, Goguet de la Salmoniere
686   YO, Aman K, Kato-Maeda M, and Small PM. 2004. Functional and evolutionary
687   genomics of *Mycobacterium tuberculosis:* insights from genomic deletions in 100
688   strains. *Proceedings of the National Academy of Sciences of the United States of*
689   *America* 101:4865-4870. 10.1073/pnas.0305634101
690 Van der Auwera GA, Carneiro M, Hartl C, Poplin R, del Angel G, Levy-Moonshine

691   A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K, Altshuler

692   D, Gabriel S and DePristo M. 2013. From FastQ Data to High-Confidence Variant Calls:
693 The Genome Analysis Toolkit Best Practices Pipeline. *Current*

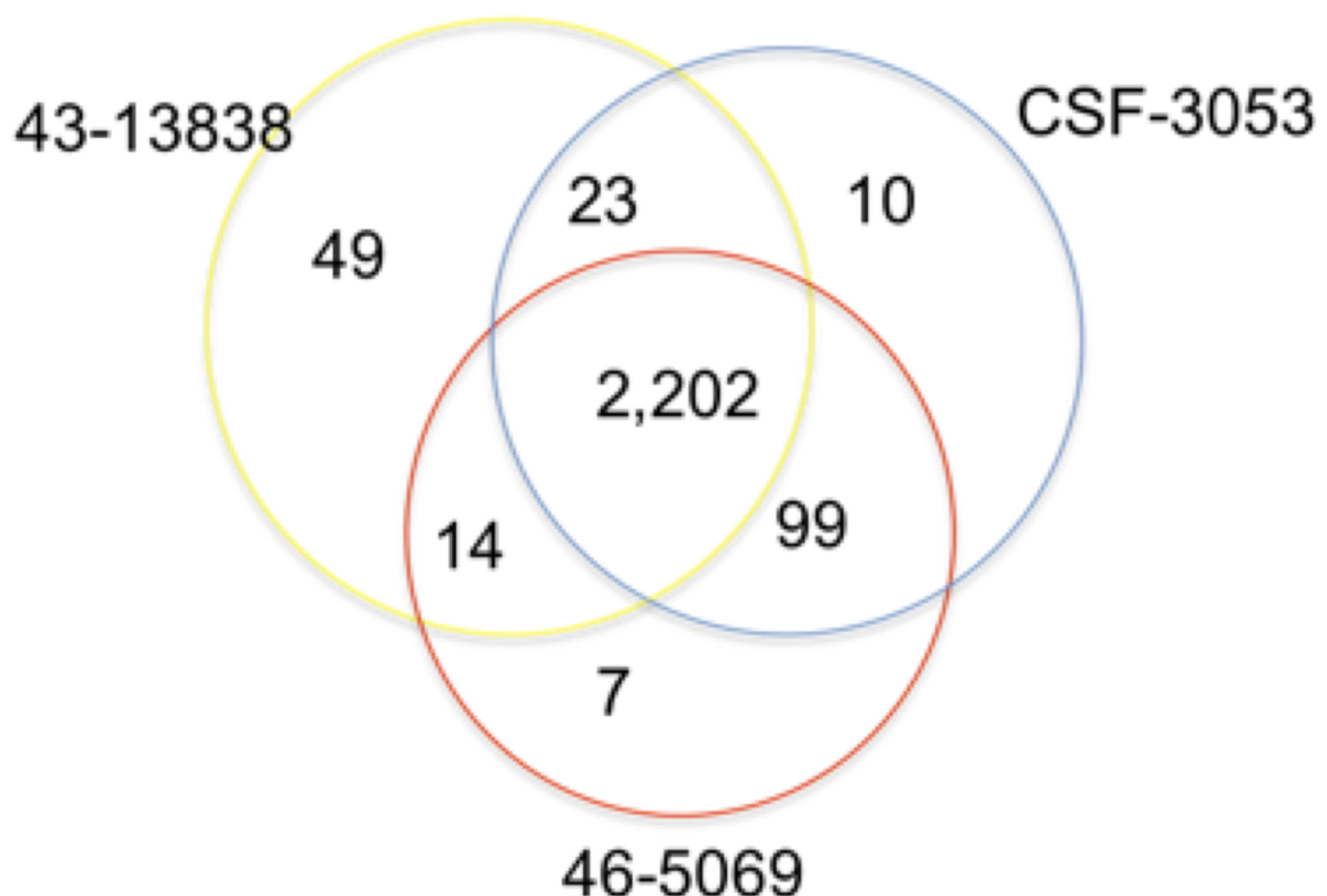
694   *Protocols in Bioinformatics* 43:11.10.1-11.10.33

695
696 van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, Hermans P, Martin
697   C, McAdam R, Shinnick TM, and Small PM. 1993. Strain identification of
698   *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a
699   standardized methodology. *Journal of Clinical Microbiology* 31:406-409.

700 Velmurugan K, Chen B, Miller JL, Azogue S, Gurses S, Hsu T, Glickman M, Jacobs WR, Jr.,
701   Porcelli SA, and Briken V. 2007. *Mycobacterium tuberculosis nuoG* is a virulence gene
702   that inhibits apoptosis of infected host cells. *PLoS Pathogens* 3:e110.
703   10.1371/journal.ppat.0030110
704 Wang XD, Gu J, Wang T, Bi LJ, Zhang ZP, Cui ZQ, Wei HP, Deng JY, and Zhang XE. 2011.
705   Comparative analysis of mycobacterial NADH pyrophosphatase isoforms reveals a
706   novel mechanism for isoniazid and ethionamide inactivation. *Molecular*
707   *Microbiology* 82:1375-1391. 10.1111/j.1365-2958.2011.07892.x
708 Zheng H, Lu L, Wang B, Pu S, Zhang X, Zhu G, Shi W, Zhang L, Wang H, Wang S, Zhao G, and
709   Zhang Y. 2008. Genetic basis of virulence attenuation revealed by comparative
710   genomic analysis of *Mycobacterium tuberculosis* strain H37Ra versus H37Rv. *PloS*
711   *One* 3:e2375. 10.1371/journal.pone.0002375
712 Zumla A, George A, Sharma V, Herbert RH, Baroness Masham of I, Oxley A, and Oliver M.
713   2015. The WHO 2014 global tuberculosis report--further to go. *The Lancet Global*
714   *Health* 3:e10-12. 10.1016/S2214-109X(14)70361-4
715
716
717

1

Distribution of single nucleotide polymorphisms in isolates CSF3053, 46-5069 and 43-13838.

Venn diagram showing the distribution of the single nucleotide polymorphisms (SNPs) observed in isolates CSF-3053 (blue), 46-5069(red) and 43-13838(yellow). CSF3053, 46-5069 and 43-13838 have 10, 7 and 49 unique SNPs respectively. 43-13838 and CSF3053 have 23 SNPs in common, CSF3053 and 46-5069 have 99 SNPs in common, while 43-13838 and 46-5069 have 14 SNPs in common. 2,202 SNPs are common to all isolates.

Statistics of whole genome sequencing, genome assembly and annotation.

1   Tables and captions

2   **Table 1: Statistics of whole genome sequencing, genome assembly and annotation**

Isolate	Total reads	% of reads mapped to Reference	% of Reference covered	Number of contigs	*N50*	Fold coverage of positions in the genome	GC content (%)	Number of predicted Genes	No. of predicted RNA genes	No. of predicted pseudo genes
CSF-3053	50,004,564	99.96	99.78	159	69,028	1329.0	65.5	4153	48	62
46-5069	44,478,206	98.67	99.82	173	63,852	920	65.5	4159	48	63
43-13838	40,767,970	98.69	99.80	177	63,019	920	65.5	4150	48	67

Gross

15   statistics of the whole genome sequence data, mapping of reads, assembly of draft genome and annotation for isolates CSF-3053, 46-5069 and 43-

16   13838. Length of reference genome (*M. tuberculosis* H37Rv, NC_000962.3) is 4,411,532 base pairs , GC: guanine/ cytocine.

Regions of deletion common to isolates CSF-3053, 46-5069 and 43-13838.

Regions of deletion and affected open reading frames found in isolates CSF-3053, 46-5069 and 43-13838. All regions were confirmed by PCR reaction as described in methods.

2   **Table 2: Regions of deletion common to isolates CSF-3053, 46-5069 and 43-13838**

Region in reference genome (H37Rv, NC_000962.3)	Length	Region of difference	Open reading frame (ORF) affected
1718912-1721213	2302	RD147c [57]	*Rv1526c*
			Rv1525 (wbbL2)
			Rv1526c
3501225-3501723	499	This study	*Rv3135*
4092082-4092921	840	RD239[57]	*Rv3651*

6   Regions of deletion and affected open reading frames found in isolates CSF-3053, 46-5069 and 43-13838. All regions were confirmed

7   by PCR reaction as described in methods.

Common SNPs found in drug resistance related genes in isolates CSF-3053, 46-5069 and 43-13838.

The reference genome positions, nucleotide change, amino acid change and effect of single nucleotide polymorphisms in drug resistance related genes that are common to isolates CSF3053, 46-5069 and 43-13838. The protein variation was determined by Protein Variation Effect Analyzer (PROVEAN), a web based protein variation analysis tool (Choi et al. 2012)

1   **Table 3: Common SNPs found in drug resistance related genes in isolates CSF-3053, 46-5069 and 43-13838**

Position in reference genome (H37Rv, NC_000962.3)	Nucleotide change	Amino acid change	Protein variation effect	Gene	Associated drug	References
6112	G>C	Met291Ile	Deleterious	*gyrB*	Quinolones	(Guillemin et al. 1998)
7362	G>C	Glu21Gln	Neutral	*gyrA*	Quinolones	(Guillemin et al. 1998)
7585	G>C	Ser95Thr	Neutral	*gyrA*	Quinolones	(Guillemin et al. 1998; Kapur et al. 1995)
8452	C>T	Ala384Val	Deleterious	*gyrA*	Quinolones	(Guillemin et al. 1998)
9143	T>C	Ile614Ile		*gyrA*	Quinolones	(Guillemin et al. 1998)
9260	G>C	Leu653Leu		*gyrA*	Quinolones	(Guillemin et

						al. 1998)
9304	G>A	Gly668Asp (N)	Neutral	*gyrA*	Quinolones	(Guillemin et al. 1998)
412280	T>G	His481Gln	Neutral	*iniA*	Ethambutol	(Ramaswamy et al. 2003)
575368	T>C	Asp7Asp		*Rv0486*	Isoniazid/Ethionamide	(Projahn et al. 2011)
763031	T>C	Ala1081Ala		*rpoB*	Rifampicin	(Taniguchi et al. 1996)
763531	G>C	Pro54Pro		*rpoC*	Rifampicin	(Comas et al. 2012)
763884	C>T	Ala172Val	Neutral	*rpoC*	Rifampicin	(Comas et al. 2012)
763886	C>A	Arg173Arg		*rpoC*	Rifampicin	(Comas et al. 2012)
1406312	A>G	His343His		*Rv1258c*	Streptomycin	(Siddiqi et al. 2004)
1417019	C>T	Cys110Tyr	Deleterious	*embR*	Ethambutol	(Ramaswam

						y et al. 2000)
1674162	C>T	Gly241Gly		*fabG1*	Isoniazid	(Lavender et al. 2005)
1792777	T>C	Ile322Val	Neutral	*Rv1592c*	Isoniazid	(Ramaswam y et al. 2003)
1792778	T>C	Glu321Glu		*Rv1592c*	Isoniazid	(Ramaswam y et al. 2003)
2154724	C>A	Arg463Leu	Neutral	*katG*	Isoniazid	(Heym et al. 1995)
2518132	C>T	Thr6Thr		*kasA*	Isoniazid	(Lee et al. 1999)
2519048	G>A	Gly312Ser	Neutral	*kasA*	Isoniazid	(Lee et al. 1999)
2521342	T>C	Asp200Asp		*accD6*	Isoniazid	(Ramaswam y et al. 2003)
3154414	A>G	Ile73Thr	Neutral	*efpA*	Isoniazid	(Ramaswam y et al. 2003)
3571834	T>G	Gln237Pro	Neutral	*nudC*	Isoniazid/Ethionamide	(Wang et al.

						2011)
3647041	A>G	Ser257Pro	Neutral	*rmlD*	Ethambutol	(Ramaswam y et al. 2000)
3647591	A>G	Asn73Asn		*rmlD*	Ethambutol	(Ramaswam y et al. 2000)
4049254	G>A	Leu243Leu		*folP1*	Para-aminosalicylic acid	(Mathys et al. 2009)
4240671	C>T	Thr270Ile	Neutral	*embC*	Ethambutol	(Ramaswam y et al. 2000)
4241042	A>G	Asn394Asp	Deleterious	*embC*	Ethambutol	(Ramaswam y et al. 2000)
4242643	C>T	Arg927Arg		*embC*	Ethambutol	(Ramaswam y et al. 2000)
4243580	G>A	Val116Val		*embA*	Ethambutol	(Telenti et al. 1997)
4244420	G>C	Val396Val		*embA*	Ethambutol	(Telenti et al. 1997)
4245969	C>T	Pro913Ser	Deleterious	*embA*	Ethambutol	(Ramaswam

						y et al. 2000; Telenti et al. 1997)
4247578	G>A	Leu355Leu		*embB*	Ethambutol	(Telenti et al. 1997)
4247646	A>C	Glu378Ala	Neutral	*embB*	Ethambutol	(Telenti et al. 1997)
4407588	T>C	Ala205Ala		*rsmG*	Streptomycin	(Okamoto et al. 2007)
4407873	C>A	Val110Val		*rsmG*	Streptomycin	(Okamoto et al. 2007)

2   The reference genome positions, nucleotide change, amino acid change and effect of single nucleotide polymorphisms in drug

3   resistance related genes that are common to isolates CSF3053, 46-5069 and 43-13838. The protein variation was determined by

4   Protein Variation Effect Analyzer (PROVEAN), a web based protein variation analysis tool (Choi et al. 2012)