



Swansea University  
Prifysgol Abertawe



## Cronfa - Swansea University Open Access Repository

---

This is an author produced version of a paper published in:

*Journal of Clinical Microbiology*

Cronfa URL for this paper:

<http://cronfa.swan.ac.uk/Record/cronfa50080>

---

### Paper:

Jones, R., Harris, L., Morgan, S., Ruddy, M., Perry, M., Williams, R., Humphrey, T., Temple, M. & Davies, A. (2019). Phylogenetic analysis of Mycobacterium tuberculosis strains in Wales using core genome MLST to analyse whole genome sequencing data. *Journal of Clinical Microbiology*

<http://dx.doi.org/10.1128/JCM.02025-18>

---

This item is brought to you by Swansea University. Any person downloading material is agreeing to abide by the terms of the repository licence. Copies of full text items may be used or reproduced in any format or medium, without prior permission for personal research or study, educational or non-commercial purposes only. The copyright for any work remains with the original author unless otherwise specified. The full-text must not be sold in any format or medium without the formal permission of the copyright holder.

Permission for multiple reproductions should be obtained from the original author.

Authors are personally responsible for adhering to copyright and publisher restrictions when uploading content to the repository.

<http://www.swansea.ac.uk/library/researchsupport/ris-support/>

1 **Phylogenetic analysis of *Mycobacterium tuberculosis* strains in Wales using core genome MLST**  
2 **to analyse whole genome sequencing data**

3

4 **Running title:** Phylogenetics of *M. tuberculosis* by cgMLST

5

6 **R.C. Jones<sup>a</sup>, L.G. Harris<sup>a</sup>, S. Morgan<sup>b</sup>, M.C.Ruddy<sup>c</sup>, M. Perry<sup>c</sup>, R. Williams<sup>c</sup>, T. Humphrey<sup>a</sup>, M.**  
7 **Temple<sup>b</sup>, A.P.Davies<sup>a,d#</sup>**

8

9 *Keywords: Mycobacterium tuberculosis; phylogenetics; whole genome sequencing; Wales; UK*

10

11 <sup>a</sup> Swansea University Medical School, Institute of Life Science, Swansea University, Swansea,  
12 Wales, UK.

13 <sup>b</sup> Health Protection Division (Mid and West Wales), Public Health Wales, Swansea, Wales, UK.

14 <sup>c</sup> Wales Centre for Mycobacteriology, Llandough Hospital, Cardiff, Wales, UK

15 <sup>d</sup> Public Health Wales Microbiology Swansea, Wales, UK

16

17 #Corresponding author: Angharad Davies angharad.p.davies@swansea.ac.uk

18

19

20 **Abstract**

21

22 Inability to standardize the bioinformatic data produced by whole genome sequencing (WGS) has  
23 been a barrier to its widespread use in tuberculosis phylogenetics. The aim of this study was to  
24 carry out a phylogenetic analysis of tuberculosis in Wales, using Ridom SeqSphere software for  
25 core genome MLST (cgMLST) analysis of whole genome sequencing data. The phylogenetics of  
26 tuberculosis in Wales has not previously been studied. Sixty-six *Mycobacterium tuberculosis*  
27 isolates (including 42 outbreak-associated isolates) from South Wales were sequenced using an  
28 Illumina platform. Isolates were assigned to Principal Genetic Groups, Single Nucleotide  
29 Polymorphism (SNP) cluster groups, lineages and sub-lineages using SNP-calling protocols. WGS  
30 data were submitted to the Ridom SeqSphere software for cgMLST analysis and analysed  
31 alongside 179 previously lineage-defined isolates. The dataset was dominated by the Euro-  
32 American lineage, with the sub-lineage composition being dominated by T, X and Haarlem family  
33 strains. The cgMLST analysis successfully assigned 58 isolates to major lineages and results were  
34 consistent with those obtained by traditional SNP mapping methods. In addition, the cgMLST  
35 scheme was used to resolve an outbreak of tuberculosis occurring in the region. This study  
36 supports the use of a cgMLST method for standardized phylogenetic assignment of tuberculosis  
37 isolates and for outbreak resolution, and provides the first insight into Welsh tuberculosis  
38 phylogenetics, identifying the presence of the Haarlem sub-lineage commonly associated with  
39 virulent traits.

40

## 41 Introduction

42 Within the species *Mycobacterium tuberculosis* seven major lineages have been recognised  
43 globally [1, 2], with different characteristics in terms of evolutionary status, transmissibility, drug  
44 resistance, host interaction, latency and vaccine efficacy [3]. Sub-lineages also show variations in  
45 virulence and pathogenicity [4]: in particular, lineage 2 (East Asian) and lineage 4 (Euro-American)  
46 contain strains, such as the Beijing and Haarlem genotypes respectively, which are notorious for  
47 their association with tuberculosis outbreaks and are over-represented amongst drug resistant  
48 cases [5, 6].

49  
50 Traditional PCR-based typing methods, such as MIRU-VNTR profiling and spoligotyping, have  
51 allowed the classification of isolates into phylogeographically related clades and families, and led  
52 to the development of readily available databases such as SpolDB4 [7, 8] and MIRU-VNTRplus [9].  
53 Two other typing methods that have been developed with results correlating with internationally  
54 recognised spoligotype families are the Principal Genetic Groupings (PGG) and SNP cluster  
55 grouping (SCG). The PGGs classifies isolates into one of three groups based on non-synonymous  
56 variants at the *katG* and *gyrA* genes [10]. SCG classifies isolates into six phylogenetically distinct  
57 groups and a further five sub-groups based on the nucleotides present at nine specific loci in the  
58 H37Rv reference genome [11, 12].

59  
60 With the advent of whole genome sequencing (WGS), comparative analysis has led to the use of  
61 single nucleotide polymorphisms (SNPs) as robust genetic markers for phylogenetic assignment [2,  
62 7]. SNPs are reliable and phylogenetically informative markers, since the low sequence variation  
63 and lack of horizontal gene transfer in *M. tuberculosis* makes independent recurrent mutations  
64 unlikely [7]. However, the lack of WGS data standardisation has been one of the barriers to its  
65 widespread usage [13, 14]. Coll *et al.* [15] developed a robust SNP barcode method analysing 60

66 loci, capable of assigning *M. tuberculosis* isolates into major lineages and sub-lineages. The  
67 method has a higher level of resolution than PGG and SCG cluster grouping, provides correlation  
68 with spoligotype families, and can be compared with a globally established database [15]. The  
69 development of WGS gene-by-gene MLST methods and software such as Ridom SeqSphere [16]  
70 has resulted in a more standardised and user-friendly approach than traditional WGS SNP mapping  
71 for resolving and understanding outbreaks [14, 17, 18]. Ridom SeqSphere allows isolate sequences  
72 to be aligned and compared in a standardised manner using a globally-defined core genome MLST  
73 (cgMLST) scheme [13, 16, 18]. To date, although this method has been used for providing clinical  
74 resolution of tuberculosis outbreaks [13], it has not been used to analyse the phylogenetic  
75 composition of a *M. tuberculosis* isolate dataset.

76

77 The phylogenetic diversity of strains of *M. tuberculosis* in Wales has not previously been studied.  
78 One aim of this work was to use for the first time the gene-by-gene based core genome MLST  
79 (cgMLST) method, PGG, SCG, and SNP bar-coding to phylogenetically analyse 66 Welsh *M.*  
80 *tuberculosis* isolates, assign them to phylogenetic groups, lineages and sub-lineages, and carry out  
81 a comparison of the different methods. Identifying the presence of strains such as Haarlem and  
82 Beijing, which are associated with outbreaks and resistance would be of interest to public health  
83 and outbreak control organisations in Wales, the UK and further afield, and give insight into the  
84 diversity of tuberculosis within Wales.

85

86 cgMLST was also used to study a set of isolates from one particular outbreak of tuberculosis in  
87 south Wales in detail. This outbreak came to the attention of Public Health Wales in 2006. At that  
88 time the outbreak involved 8 cases with cultured isolates and appeared to be circulating amongst  
89 individuals who frequented five local public houses within an area, with one public house having  
90 connections to several cases in the outbreak. The index case was the landlord of that public house

91 and at the time of that diagnosis in 2004, contact tracing of close contacts and the pub's regular  
92 customers had been carried out promptly and detected no other cases. The outbreak sparked a  
93 review by Public Health Wales of tuberculosis case records in the area. From 2006-2011 a further  
94 five cases with clinical isolates were reported, making a total of 13 reported isolate-confirmed  
95 cases in the area since 2004. Two were an estranged husband and wife pair. All the isolates were  
96 fully susceptible to all first line anti-tuberculous chemotherapy.

97

98

99 **Materials and Methods**

100 **Isolates**

101 DNA from 66 *M. tuberculosis* isolates collected between 2004 and 2011 were obtained from the  
102 Wales Centre for Mycobacteriology Cardiff, UK. Forty-two of the isolates were from 3 separate  
103 tuberculosis outbreaks in the south west area of Wales according to both MIRU-VNTR typing and  
104 epidemiological investigations (isolate prefixes LL, NPT, TH or GO), and the remaining 24 were  
105 randomly selected endemic isolates (prefix BK). Outbreak isolates prefixed NPT  
106 were those from one particular public house-related outbreak of tuberculosis which was studied in detail,  
107 as outlined in the Introduction.

108

109 **Epidemiological investigation**

110 Epidemiological information was obtained from face-to-face interviews with a nurse from the  
111 original PHW contact tracing investigation team and from documents produced during the  
112 outbreak investigation.

113

114 **Sequencing and assembly**

115 The genomic DNA was sequenced using Nextera XT library preparation kits (Version 3, Illumina)  
116 and a MiSeq benchtop sequencer (Illumina, San Diego, CA, USA), with paired-end reads quality  
117 filtered with the Trimmomatic tool software version 0.32 (Usadellab, Germany) using a sliding  
118 window approach of 5 bases and a quality score of Q20. The resulting contigs/genomes were  
119 assembled using SPAdes genome assembler version 3.9.0 [19]. K-mers used for SPAdes were 33,  
120 55, 77.99 and 127. The sequence read archive (SRA) sequences for 179 lineage-defined isolates  
121 (NCBI) previously published [1] were also assembled using the SPAdes genome assembler.

122

123 **Core genome MLST analysis (cgMLST) and phylogenetic assignment**

124 Assembled genomes were uploaded onto the Ridom SeqSphere software version 4.1.9 (Ridom;  
125 Münster, Germany). Each isolate sequence was aligned to the Ridom SeqSphere *M. tuberculosis*  
126 cgMLST scheme of 2891 core genes (GenBank accession number NC\_000962.3), previously defined  
127 for alignment and subsequent genomic analysis [14,18]. Successful alignments to the cgMLST were  
128 defined as “good targets” by the Ridom SeqSphere software, and full cgMLST analysis was carried  
129 out on isolate sequences that conferred >90% “good targets”. The cgMLST scheme was also used  
130 to compare the sequenced Welsh isolates and 179 isolates previously lineage-defined by Comas *et*  
131 *al* [1]. The 179 isolates selected from [1] were those whose genomes also exceeded the 90%  
132 quality threshold under the Ridom SeqSphere parameters. The resulting phylogeny comparison  
133 was made using an Unweighted Pair Group Method with Arithmetic Mean (UPGMA) tree  
134 produced by the Ridom SeqSphere, and further annotated and modified using iTol version 4 (  
135 <https://itol.embl.de>) [20]. The genome of *Mycobacterium canetti*, as the ancestral member of the  
136 *M. tuberculosis* complex, was used to root the tree.

137

138 **WGS SNP bar-coding and sub-lineage genotyping**

139 Isolates were aligned to the H37Rv reference genome using Burrows-Wheeler Alignment (BWA,  
140 version 0.7.17) [21]. SAMtools version 1.3.1 [22] was then used to call SNPs from each of 60  
141 designated loci previously described [15] (with the omission of two *M. bovis* loci). Thus the  
142 isolates based on the SNPs pattern (SNP barcode) at the designated loci were split into one of the  
143 phylogeographically related groups: Lineage 1 (Indo-Oceanic), Lineage 2 (East Asian), Lineage 3  
144 (East African-Indian), Lineage 4 (Euro-American), Lineage 5 (West Africa 1), Lineage 6 (West Africa  
145 2), or Lineage 7 (Horn of Africa) [7, 15].

146

147 Each *M. tuberculosis* lineage determined by SNP mapping was also divided into one of the  
148 following sub-lineages: Beijing [23], Latin American Mediterranean (LAM) [24], Haarlem [25] or X  
149 family [24]. SNPs were initially identified through extraction of relevant gene sequences from  
150 each isolate using the sequence extraction application within Ridom SeqSphere and detected  
151 manually using BioEdit. Concatenated SNPs were then used to produce a phylogenetic UPGMA  
152 tree using the iTol software, and isolates assigned to one of the sub-lineage genotypes listed  
153 above.

154

#### 155 **Principal Genetic Grouping (PGG) and SNP Cluster Grouping (SCG)**

156 Gene sequences for *gyrA* and *katG* were extracted from the WGS of each isolate using Ridom  
157 SeqSphere and analysed manually using BioEdit to identify the presence of PGG-defining amino  
158 acids at codons 95 and 493, the PGG informative sites within genes *gyrA* and *katG* [10]. Based on  
159 the composition of amino acids at these loci, each isolate was assigned a PGG group [26]. For SCG  
160 analysis, sequences were aligned to the H37Rv reference genome using BWA. SAMtools was then  
161 used to call SNPs from the previously defined nine specific loci [12] and each isolate then assigned  
162 to a SNP cluster group. Phylogenetic analysis was only carried out on isolates with each of the nine  
163 loci present (31 isolates).

164

165 **Results**

166 N50 and number of contigs for each assembled genome are shown in Supplementary Table S1.

167

168 **cgMLST association**

169 Fifty eight of the 66 isolates had a sequence quality sufficient for cgMLST analysis and were  
170 incorporated into a phylogeny that also included the 179 lineage-defined isolates [1]. The resulting  
171 tree shows the Welsh and lineage-defined isolates clustered into lineages 1 (n=1), 2 (n=3) and 4  
172 (n=53) (Figure 1). Lineages 3, 5, 6 and 7 are not shown as none of the Welsh isolates were assigned  
173 to them. All but one outbreak-associated isolate (LL9) clustered with the lineage 4 isolates, while  
174 the endemic isolates showed more lineage diversity.

175

176 **Phylogenetic composition using SNP bar-coding and sub-lineage genotyping**

177 SNP bar-coding was carried out on the 59 Welsh isolates that had >90% sequence data as required  
178 for the 60 loci SNP barcode analysis. The results were consistent with those from cgMLST  
179 association. Lineage 4 (Euro-American) dominated the dataset with 55 isolates (Figure 2), and all  
180 but one outbreak-associated isolate clustered with this lineage. Fourteen of the 55 lineage 4  
181 isolates were of the Haarlem sub-lineage, and of the 18 T family isolates, 13 showed a clonal  
182 pattern across the 60 SNPs, with 10 of these from the same recognised outbreak. Twelve of the 16  
183 X family could be split into three clonally-related clusters correlating to that seen in Figure 2b, and  
184 three lineage 2 Beijing strains were identified. The T family sub-lineage dominated the outbreak  
185 isolates (39%), followed by the Haarlem sub-lineage (33%) and the X family (27%) respectively.  
186 Table 1 shows a direct comparison between the cgMLST and SNP results, indicating correlation at  
187 the lineage level for each Welsh isolate.

188

189 **PGG and SCG analysis**

190 Of the 66 isolates sequenced, fifty-seven could be assigned to a PGG based on sequence data as  
191 shown in Figure 3a. Four isolates clustered within PGG1, 31 within PGG2, and 22 within PGG3,  
192 along with the H37Rv genome. When compared to the sub-lineage data, the Haarlem, X family and  
193 LAM sub-lineage isolates grouped with PGG2 and the T family and H37Rv-like isolates with PGG3.  
194 All lineage 1 and 2 isolates were associated with PGG1. Fifty-six of the original 66 isolates could be  
195 assigned confidently to an SCG based on the sequence data provided. The SCG results identified  
196 two predominant SCGs, SCG-6a and SCG-3b with 16 and 15 isolates clustering to these sub-groups  
197 respectively (Figure 3b). Other sub-groups present were SCG-4 (8 isolates), SCG-3c (7 isolates),  
198 SCG-6b (4 isolates), SCG-5 (3 isolates), SCG-2 (2 isolates) and SCG-1 (1 isolate). Nine isolates were  
199 excluded as they did not yield sequence data for all nine loci and SCG-3a was not represented in  
200 the dataset. The SCG phylogeny split into two clear clades, with clade 2 being more diverse than  
201 clade 1. When PGG results were compared with SCG results, it was found that clade 1 contained  
202 all the PGG3 isolates and clade 2 contained all PGG1 and PGG2 isolates (Figure 3b). The PGG2  
203 isolates also divided into four different SCG groups. Within clade 2, SCG-3c and SCG-4 shared a  
204 closer relationship with each other than they did with isolates of SCG-3b and SCG-5, and vice  
205 versa.

206

#### 207 **NPT outbreak isolate analysis**

208 All the NPT-designated outbreak isolates clustered as Euro-American T family isolates, except for  
209 NPTB6 (Figure 2). In addition, a further 3 three background isolates (BK1, BK2 and BK3) also  
210 clustered clonally as T family isolates and were included in further downstream analysis (Figure 2).  
211 NPTB6 did not cluster within the same T family sub-lineage, but clustered with 6 X family sub-  
212 lineage isolates. This was evidence that NPTB6 had been wrongly included within this outbreak  
213 cluster and was unrelated. For further outbreak analysis, the 3 additional T family background  
214 cases were included with the NPT isolates when analysed with cgMLST.

215 cgMLST analysis revealed that there were in fact 8 distinct isolates within the T family group,  
216 including the existence of 2 clusters (Figure 4). The clusters defined by cgMLST consisted of one  
217 containing 9 isolates (Outbreak 1) and one containing 2 isolates (Outbreak 2; the estranged  
218 husband and wife). In Outbreak 1 there were 8 NPT isolates and one background isolate,  
219 previously thought of as an unrelated case. NPTA3 showed 16 allelic differences from its closest  
220 relative (NPTA7) and thus according to the definition of no more than 12 allelic differences [13,14]  
221 could not be directly linked to either outbreak. Five other isolates showed no evidence of being  
222 directly linked with any other isolate within the dataset: these included three NPT isolates (NPTB2,  
223 NPTB5 and NPTB6), and two background ones (BK1 and BK3). The data indicated that NPTA7 was  
224 the source case. This case, diagnosed with pulmonary tuberculosis in 2007, was known to other a  
225 number of the other cases as a regular at the public house, although he denied this. The cgMLST  
226 results supported the epidemiological evidence that he was associated with the public house.

227

## 228 Discussion

229 This study has provided the first insight into the phylogenetic diversity of *M. tuberculosis* isolates  
230 from Wales using cgMLST. In addition it is one of the first independent confirmatory studies of  
231 Kohl *et al*'s cgMLST scheme. Gene-by-gene MLST methods have previously been shown to be  
232 useful in clinical outbreak resolution and epidemiological investigations of human pathogens such  
233 as MRSA and *Campylobacter*, as well as *M. tuberculosis* itself [17, 18]. Specifically, the Ridom  
234 SeqSphere gene-by-gene cgMLST scheme has been used previously to look at tuberculosis  
235 outbreaks [13, 18], and consists of a portable, standardised database platform for use with WGS  
236 data in tuberculosis research. However, the method has not been used previously for classification  
237 of *M. tuberculosis* into well-defined phylogenetic lineages. This study provided for the first time a  
238 snapshot of the tuberculosis phylogenetics across a geographical area based on cgMLST, in  
239 comparison with SNP calling methods. In this study, the resulting cgMLST phylogenetic tree  
240 contained all seven major *M. tuberculosis* sub-lineages and broadly matched that seen using SNP  
241 mapping-based methods [1, 27]. Of the 66 isolates WGS, 58 were successfully analysed by cgMLST  
242 in conjunction with 179 lineage-defined isolates [1], with lineage 4, the Euro-American lineage  
243 dominating the collection. Lineage 1 and 2 isolates were also identified, but in much lower  
244 numbers. Consistent with Comas *et al.* [1], lineages 2 and 3 isolates shared a closer relationship  
245 with each other than with lineage 4 isolates. Hence, despite using a different set of genomic data,  
246 the evolutionary positions of each lineage according to cgMLST was consistent with other studies  
247 that used in-house SNP mapping pipelines for the construction of their phylogenies [1, 27, 28].

248

249 According to the SNP barcoding and sub-genotyping methods, which correlated with the cgMLST  
250 results, the dataset contained a diverse collection of Euro-American sub-lineages, which were not  
251 dominated by a single sub-lineage, as T family, X family and the Haarlem family made up a large  
252 proportion of the lineage 4 dataset, with the Haarlem isolates being particularly prevalent within

253 the outbreak-assigned cases. The proportion of Euro-American lineage isolates here is similar to  
254 Public Health England data for TB cases in indigenous people across the whole of the UK and  
255 Ireland [29]. This study also identified 2% of the isolates as lineage 1 and 6% as lineage 2, again  
256 correlating with the data for the indigenous population of the UK [29] and Ireland [30, 31]. The  
257 discovery of numerous Haarlem sub-lineage strains, and some Beijing strains, was an interesting  
258 finding.

259

260 The PGG results correlated well with the lineage groupings, as 31 of the Welsh isolates were PGG2  
261 or PGG3, which have previously been associated with the Euro-American lineage, whilst PGG1 is  
262 associated with lineages 1, 2 and 3 [7]. The SCG results revealed a predominance of SCG-3 and  
263 SCG-6 isolates, with SCG-3b and SCG-6a being the most prominent. Unlike for PGG, the SCG  
264 analysis highlighted a large degree of divergence within the Euro-American lineage, consistent  
265 with the diversity seen in the SNP barcode result. Such an association was expected as SCGs have  
266 previously been shown to assign themselves with the SNP bar-coding and sub-lineage groupings  
267 [7, 11].

268 Phylogenetic analysis confirmed that all the apparent NPT outbreak isolates except NPTB6, were  
269 clustered within the same sub-lineage, the Euro-American T family. In addition, the SNP barcode  
270 method identified three further apparently unrelated local isolates that clustered within this  
271 phylogeny; indicating that phylogenetic characterisation may be useful in tuberculosis outbreak  
272 investigation.

273 Through the use of cgMLST, the relationship between the NPT outbreak isolates was resolved, and  
274 two clusters/outbreaks were confirmed. The cgMLST analysis also confirmed that the cases in  
275 Outbreak 1 were directly linked to the public house, as assumed by initial contact tracing team.  
276 However a number of cases, including the estranged husband and wife pair, were unrelated,

277 serving as a reminder that TB remains endemic in Wales and cases occurring within a small area  
278 are not necessarily related. Such results could be used as a basis to support targeted outbreak  
279 control interventions around the public house, and the identification of NPTA7 (who denied  
280 frequenting the public house, contradicting the evidence of other cases) as the source case.

281 SNP barcoding provides a very high level of resolution, is more established in terms of providing  
282 sub-lineage assignments and provides correlation with spoligotyping. However, it requires  
283 bioinformatic expertise and is difficult to standardise as it is not linked to a global database. In  
284 addition, the SNP barcode used here is based solely on a set of markers (15) and so cannot provide  
285 understanding of individual relationships within an outbreak, restricting its use to phylogenetics.

286 In comparison, cgMLST is a relatively new method. However it has the advantage of being a  
287 simpler, standardised method for analysing large amounts of genomic data which are easily  
288 uploaded to a global database for analysis using the user-friendly Ridom SeqSphere software, which  
289 could facilitate the use of genomics for tuberculosis surveillance. The results of cgMLST analysis  
290 were consistent with those obtained by traditional SNP mapping methods. Although cgMLST is yet  
291 to be developed to a level whereby isolates can be confidently assigned to a phylogenetic sub-  
292 lineage, this study provides evidence that, at least at lineage level, the phylogenetic associations  
293 made using cgMLST correlate with those from SNP barcoding. This work supports the use of  
294 cgMLST for standardized phylogenetic assignment of tuberculosis isolates, in addition to its use for  
295 delineating clinical outbreaks (13, 18).

## 296 **Acknowledgements**

297 This work was funded by St. David's Medical Foundation & Coleg Cenedlaethol Cymraeg funding .  
298 The funders had no role in study design, data collection and interpretation, or the decision to  
299 submit the work for publication.

300 **References**

301

302 1. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg  
303 S, Thwaites G, Yeboah-Manu D, Bothamley G, Mei J, Wei L, Bentley S, Harris SR, Niemann  
304 S, Diel R, Aseffa A, Gao Q, Young D, Gagneux S. 2013. Out-of-Africa migration and Neolithic  
305 co expansion of *Mycobacterium tuberculosis* with modern humans. *Nature Genet* 45:1176-  
306 1182.

307

308 2. Gagneux S, Deriemer K, Van T, Kato-Maeda M, De Jong BC, Narayanan S, Nicol M, Niemann  
309 S, Kremer K, Gutierrez MC, Hilty M, Hopewell PC, Small PM. 2006. Variable host–pathogen  
310 compatibility in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci U S A.* 103:2869-2873.

311

312 3. Thwaites G, Caws M, Chau TTH, D’Sa A, Lan NT, Huyen MNT, Gagneux S, Anh PT, Tho  
313 DQ, Torok E, Nhu NT, Duyen NT, Duy PM, Richenberg J, Simmons C, Hien TT, Farrar J. 2008.  
314 Relationship between *Mycobacterium tuberculosis* genotype and the clinical phenotype of  
315 pulmonary and meningeal tuberculosis. *J Clin Micro.* 46:1363-1368.

316

317 4. Anderson J, Jarlsberg LG, Grindsdale J, Osmond D, Kawamura M, Hopewell PC, Kato-Maeda  
318 M. 2013. Sublineages of lineage 4 (Euro-American) *Mycobacterium tuberculosis* differ in  
319 genotypic clustering. *Int J Tuberc Lung Dis.* 17:885-891.

320

321 5. Marais BJ, Victor TC, Hesseling AC, Barnard M, Jordaan A, Brittle W, Reuter H, Beyers N,  
322 van Helden PD, Warren RM, Schaaf HS. 2006. Beijing and Haarlem genotypes are  
323 overrepresented among children with drug-resistant tuberculosis in the Western Cape  
324 Province of South Africa. *J Clin Micro.* 44:3539-3543.

325

326 6. Bifani PJ, Plikaytis BB, Kapur V, Stockbauer K, Pan X, Lutfey ML, Moghazeh SL, Eisner W,  
327 Daniel TM, Kaplan MH, Crawford JT, Musser JM, Kreiswirth BN. 1996. Origin and interstate  
328 spread of a New York City multidrug-resistant *Mycobacterium tuberculosis* clone family.  
329 JAMA. 275:452-457.

330

331 7. Gagneux S. & Small PM. 2007. Global phylogeography of *Mycobacterium tuberculosis* and  
332 implications for tuberculosis product development. Lancet Infect Dis. 7:328-337.

333

334 8. Brudey K, Driscoll JR, Rigouts L, Prodinger W, Gori WM, Al-Hajj SA, Allix C, Aristimuño L,  
335 Arora J, Baumanis V, Binder L, Cafrune P, Cataldi A, Cheong S, Diel R, Ellermeier C, Evans JT,  
336 Fauville-Dufaux M, Ferdinand S, Garcia de Viedma D, Garzelli C, Gazzola L, Gomes HM,  
337 Guttierrez MC, Hawkey PM, van Helden PD, Kadival GV, Kreiswirth BN, Kremer K, Kubin M,  
338 Kulkarni SP, Liens B, Lillebaek T, Ho ML, Martin C, Martin C, Mokrousov I, Narvskaia O,  
339 Ngeow YF, Naumann L, Niemann S, Parwati I, Rahim Z, Rasolofo-Razanamparany V,  
340 Rasolonalona T, Rossetti ML, Rüsç-Gerdes S, Sajduda A, Samper S, Shemyakin IG, Singh  
341 UB, Somoskovi A, Skuce RA, van Soolingen D, Streicher EM, Suffys PN, Tortoli E, Tracevska  
342 T, Vincent V, Victor TC, Warren RM, Yap SF, Zaman K, Portaels F, Rastogi N, Sola C. 2006.  
343 *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international  
344 spoligotyping database (SpolDB4) for classification, population genetics and epidemiology.  
345 BMC Microbiol. 6:23.

346

347 9. Weniger T, Krawczyk J, Supply P, Niemann S, Harmsen D. 2010. MIRU-VNTRplus: a web tool  
348 for polyphasic genotyping of *Mycobacterium tuberculosis* complex bacteria. Nucleic Acids  
349 Res. 38:W326-31.

350

351 10. Sreevatsan S, Pan X, Stockbauer K, Connell N, Kreiswirth B, Whittam T, Musser JM. 1997.  
352 Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex  
353 indicates evolutionarily recent global dissemination. Proc Natl Acad Sci U S A. 94:9869-  
354 9874.

355

356 11. Filliol I, Motiwala AS, Cavatore M, Qi W, Hazbon MH, Bobadilla Del Valle M, Fyfe J, García-  
357 García L, Rastogi N, Sola C, Zozio T, Guerrero MI, León CI, Crabtree J, Angiuoli S, Eisenach  
358 KD, Durmaz R, Joloba ML, Rendón A, Sifuentes-Osornio J, Ponce de León A, Cave MD,  
359 Fleischmann R, Whittam TS, Alland D. 2006. Global Phylogeny of *Mycobacterium*  
360 *tuberculosis* Based on Single Nucleotide Polymorphism (SNP) Analysis: Insights into  
361 Tuberculosis Evolution, Phylogenetic Accuracy of Other DNA Fingerprinting Systems, and  
362 Recommendations for a Minimal Standard SNP Set. J Bact. 188:759-772

363

364 12. Alland D, Lacher DW, Hazbon MH, Motiwala AS, Qi W, Fleischmann RD, Whittam TS. 2007.  
365 Role of large sequence polymorphisms (LSPs) in generating genomic diversity among  
366 clinical isolates of *Mycobacterium tuberculosis* and the utility of LSPs in phylogenetic  
367 analysis. J Clin Microbiol. 45:39-46.

368

369 13. Kohl TA, Diel R, Harmsen D, Rothganger J, Walter KM, Merker M, Weniger T, Niemann S.  
370 2014. Whole-genome-based *Mycobacterium tuberculosis* surveillance: a standardized,  
371 portable, and expandable approach. J Clin Microbiol. 52:2479-2486

372

373 14. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dediccoat MJ, Eyre DW, Wilson DJ,  
374 Hawkey PM, Crook DW, Parkhill J, Harris D, Walker AS, Bowden R, Monk P, Smith EG, Peto

375 TE. 2013. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks:  
376 a retrospective observational study. *Lancet Infect Dis.* 13:137-146.  
377

378 15. Coll F, McNerney R, Guerra-Assuncao JA, Glynn JR, Perdigao, Viveiros M, Portugal I, Pain A,  
379 Martin N, Clark TG. 2014. A robust SNP barcode for typing *Mycobacterium tuberculosis*  
380 complex strains. *Nat Commun.* 5:4812.  
381

382 16. Junemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U, Kalinowski J, Mellmann A,  
383 Goesmann A, von Haeseler A, Stoye J, Harmsen D. 2013. Updating benchtop sequencing  
384 performance comparison. *Nat Biotechnol.* 31:294-6.  
385

386 17. Maiden MC, Van Rensburg MJJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND. 2013.  
387 MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat Rev Microbiol.*  
388 11:728-736.  
389

390 18. Kohl TA, Harmsen D, Rothganger J, Walker T, Diel R, Niemann S. 2018. Harmonised  
391 Genome Wide Typing of Tubercle Bacilli Using a Web-Based Gene-By-Gene Nomenclature  
392 System. *EBioMedicine* 34: 131-138.  
393

394 19. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko  
395 SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA,  
396 Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to  
397 single-cell sequencing. *J Comput Biol.* 19:455-477  
398

- 399 20. Letunic I & Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and  
400 annotation of phylogenetic and other trees. *Nucleic Acids Res.* 44(W1):W242-5  
401
- 402 21. Li H & Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler  
403 transform. *Bioinformatics.* 25(14):1754-60  
404
- 405 22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R;  
406 1000 Genome Project Data Processing Subgroup. 2009. The Sequence alignment/map  
407 (SAM) format and SAMtools. *Bioinformatics.* 25:2078-9  
408
- 409 23. Mestre O, Luo T, Dos Vultos T, Kremer K, Murray A, Namouchi A, Jackson C, Raugier J, Bifani  
410 P, Warren R, Rasolofo V, Mei J, Gao Q, Gicquel B. 2011. Phylogeny of *Mycobacterium*  
411 *tuberculosis* Beijing strains constructed from polymorphisms in genes involved in DNA  
412 replication, recombination and repair. *PLoS One.* 6:e16020.  
413
- 414 24. Comas I, Homolka S, Niemann S, Gagneux S. 2009. Genotyping of genetically monomorphic  
415 bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of  
416 current methodologies. *PLoS One.* 4:e7815.  
417
- 418 25. Cubillos-Ruiz A, Sandoval A, Ritacco V, Lopez B, Robledo J, Correa N, Hernandez-Neuta I,  
419 Zambrano MM, Del Portillo P. 2010. Genomic Signatures of the Haarlem Lineage of  
420 *Mycobacterium tuberculosis*: Implications of Strain Genetic Variation in Drug and Vaccine  
421 Development. *J Clin Microbiol.* 48:3614-3623.  
422

- 423 26. Grimes CZ, Teeter LD, Hwang L-Y, Graviss EA. 2009. Epidemiologic characterization of  
424 culture positive *Mycobacterium tuberculosis* patients by *katG-gyrA* principal genetic  
425 grouping. J Mol Diagn. 11:472-481.
- 426
- 427 27. Gagneux, S. Host–pathogen coevolution in human tuberculosis. 2012. Philos Trans Royal  
428 Soc B. 367:850-859.
- 429
- 430 28. Firdessa R, Berg S, Hailu E, Schelling E, Gumi B, Erenso G , Gadisa E, Kiros T, Habtamu M,  
431 Hussein J, Zinsstag J, Robertson BD, Ameni G, Lohan AJ, Loftus B, Comas I, Gagneux S,  
432 Tschopp R, Yamuah L, Hewinson G, Gordon SV, Young DB, Aseffa A. 2013. Mycobacterial  
433 lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia Emerg Infect Dis.  
434 19:460-463.
- 435
- 436 29. Tuberculosis in the UK 2014 report. 2014. Public Health England.
- 437
- 438 30. Fitzgibbon M, Gibbons N, Roycroft E, Jackson S, O’Donnell J, O’Flanagan D, Rogers TR.  
439 2013. A snapshot of genetic lineages of *Mycobacterium tuberculosis* in Ireland over a two-  
440 year period, 2010 and 2011. Euro Surveill. 8(3):pii:20367
- 441
- 442 31. Ojo OO, Sheehan S, Corcoran DG, Nikolayevsky V, Brown T, O’Sullivan M, O’Sullivan K,  
443 Gordon SV, Drobniewski F, Prentice MB. 2010. Molecular epidemiology of *Mycobacterium*  
444 *tuberculosis* clinical isolates in Southwest Ireland. Infection, Genetics and Evolution. 10:1110-  
445 1116.
- 446
- 447

448 **Figure Titles and Legends**

449

450 **Figure 1:** An Unweighted Pair Group Method with Arithmetic Mean tree based on the  
451 cgMLST association between 58 Welsh isolates

452

453 Unweighted Pair Group Method with Arithmetic Mean (UPGMA) tree showing the  
454 phylogeny of the 58 Welsh isolates and the lineage-defined isolates [1] which had a  
455 sequence quality sufficient for cgMLST analysis. *M. canetti* genome was used to root the  
456 tree. Lineages 3, 5, 6 and 7 are not shown as none of the Welsh isolates were assigned to  
457 them. Lineage 1 = green, Lineage 2 = yellow, Lineage 4 = pink, Welsh isolates = red, *M.*  
458 *canetti* = grey.

459

460

461 **Figure 2:** Phylogenetic analysis of 59 Welsh *M. tuberculosis* isolates

462

463 The figure shows the 59 isolates that had >90% sequence data (as required for the 60  
464 loci SNP barcode analysis) assigning the isolates to lineages and sub-lineages. A)  
465 Unweighted Pair Group Method with Arithmetic Mean (UPGMA) tree result showing  
466 SNP bar-coding results. The scale bar indicates the genetic divergence relevant to  
467 branch length and is based on units of nucleotide differences per site across 60 loci.  
468 B) Graph summarising the number of isolates representing each sub-lineage present  
469 among 59 Welsh *M. tuberculosis* isolates.

470

471

472 **Figure 3:** Neighbour joining phylogeny showing the Principal Genetic Grouping and Single  
473 Nucleotide Polymorphism Cluster Grouping profiles of 57 and 56 Welsh isolates  
474 respectively, with the reference genome H37Rv also being assigned.

475

476 A) PGG results; red: PGG1, green: PGG2, blue: PGG3. Letters refer to the amino acids  
477 present at each locus: T = Threonine, R = Arginine, L = Leucine, S = Serine. The scale bar  
478 highlights the genetic divergence relevant to branch length and is based on units of  
479 amino acid differences per site across the *gyrA* and *katG* loci. B) SCG results, where the  
480 phylogeny harbours two clades, Clade 1 and Clade 2. The PGG assigned to each isolate is  
481 shown in the right column, and X denotes isolates that could not be assigned a PGG  
482 group.

483

484

485 **Figure 4:** A minimum spanning tree of 17 cases constructed using Ridom SeqSphere  
486 software. Isolates sharing less than 12 allelic difference are classed as direct transmission  
487 events and are thus part of a clonal outbreak and are grouped accordingly into Outbreak  
488 1 and Outbreak 2.

489

490 **Table 1:** Lineage, by cgMLST and SNP analysis, of 58 sequenced isolates that had sequence  
491 quality sufficient for cgMLST analysis, showing correlation of both methods at the lineage level  
492 for each Welsh isolate.

493