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

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

R.C. Jones[a], L.G. Harris[a], S. Morgan[b], M.C.Ruddy[c], M. Perry[c], R. Williams[c], T. Humphrey[a], M. Temple[b], A.P.Davies[a,d#]

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

[a] Swansea University Medical School, Institute of Life Science, Swansea University, Swansea, Wales, UK.

[b] Health Protection Division (Mid and West Wales), Public Health Wales, Swansea, Wales, UK.

[c] Wales Centre for Mycobacteriology, Llandough Hospital, Cardiff, Wales, UK

[d] Public Health Wales Microbiology Swansea, Wales, UK

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

Abstract

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

Introduction

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

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

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

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

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

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

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

Materials and Methods

Isolates

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

Epidemiological investigation

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

Sequencing and assembly

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

Core genome MLST analysis (cgMLST) and phylogenetic assignment

Assembled genomes were uploaded onto the Ridom SeqSphere software version 4.1.9 (Ridom; Münster, Germany). Each isolate sequence was aligned to the Ridom SeqSphere *M. tuberculosis* cgMLST scheme of 2891 core genes (GenBank accession number NC_000962.3), previously defined for alignment and subsequent genomic analysis [14,18]. Successful alignments to the cgMLST were defined as "good targets" by the Ridom SeqSphere software, and full cgMLST analysis was carried out on isolate sequences that conferred >90% "good targets". The cgMLST scheme was also used to compare the sequenced Welsh isolates and 179 isolates previously lineage-defined by Comas *et al* [1]. The 179 isolates selected from [1] were those whose genomes also exceeded the 90% quality threshold under the Ridom SeqSphere parameters. The resulting phylogeny comparison was made using an Unweighted Pair Group Method with Arithmetic Mean (UPGMA) tree produced by the Ridom SeqSphere, and further annotated and modified using iTol version 4 (https://itol.embl.de) [20]. The genome of *Mycobacterium canetti*, as the ancestral member of the *M. tuberculosis* complex, was used to root the tree.

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

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

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

Principal Genetic Grouping (PGG) and SNP Cluster Grouping (SCG)

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

Results

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

cgMLST association

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

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

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

PGG and SCG analysis

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

NPT outbreak isolate analysis

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

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

Discussion

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

According to the SNP barcoding and sub-genotyping methods, which correlated with the cgMLST results, the dataset contained a diverse collection of Euro-American sub-lineages, which were not dominated by a single sub-lineage, as T family, X family and the Haarlem family made up a large 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,

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

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

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

Acknowledgements

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

20. Letunic I & Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 44(W1):W242-5

21. Li H & Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25(14):1754-60

22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. 2009. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics. 25:2078-9

23. Mestre O, Luo T, Dos Vultos T, Kremer K, Murray A, Namouchi A, Jackson C, Rauzier J, Bifani P, Warren R, Rasolofo V, Mei J, Gao Q, Gicquel B. 2011. Phylogeny of *Mycobacterium tuberculosis* Beijing strains constructed from polymorphisms in genes involved in DNA replication, recombination and repair. PLoS One. 6:e16020.

24. Comas I, Homolka S, Niemann S, Gagneux S. 2009. Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. PLoS One. 4:e7815.

25. Cubillos-Ruiz A, Sandoval A, Ritacco V, Lopez B, Robledo J, Correa N, Hernandez-Neuta I, Zambrano MM, Del Portillo P. 2010. Genomic Signatures of the Haarlem Lineage of *Mycobacterium tuberculosis*: Implications of Strain Genetic Variation in Drug and Vaccine Development. J Clin Microbiol. 48:3614-3623.

26. Grimes CZ, Teeter LD, Hwang L-Y, Graviss EA. 2009. Epidemiologic characterization of culture positive *Mycobacterium tuberculosis* patients by *katG-gyrA* principal genetic grouping. J Mol Diagn. 11:472-481.

27. Gagneux, S. Host–pathogen coevolution in human tuberculosis. 2012. Philos Trans Royal Soc B. 367:850-859.

28. Firdessa R, Berg S, Hailu E, Schelling E, Gumi B, Erenso G , Gadisa E, Kiros T, Habtamu M, Hussein J, Zinsstag J, Robertson BD, Ameni G, Lohan AJ, Loftus B, Comas I, Gagneux S, Tschopp R, Yamuah L, Hewinson G, Gordon SV, Young DB, Aseffa A. 2013. Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia Emerg Infect Dis. 19:460-463.

29. Tuberculosis in the UK 2014 report. 2014. Public Health England.

30. Fitzgibbon M, Gibbons N, Roycroft E, Jackson S, O'Donnell J, O'Flanagan D, Rogers TR. 2013. A snapshot of genetic lineages of *Mycobacterium tuberculosis* in Ireland over a two-year period, 2010 and 2011. Euro Surveill. 8(3):pii:20367

31. Ojo OO, Sheehan S, Corcoran DG, Nikolayevsky V, Brown T, O'Sullivan M, O'Sullivan K, Gordon SV, Drobniewski F, Prentice MB. 2010. Molecular epidemiology of *Mycobacterium tuberculosis* clinical isolates in Southwest Ireland. Infection, Genetics and Evolution. 10:1110-1116.

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

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

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

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

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