Resolving a clinical tuberculosis outbreak using palaeogenomic genome reconstruction methodologies

Rhys Jones[a], Marcela Sandoval Velasco[b], Llinos G Harris[a], Sue Morgan[c], Mark Temple[c], Michael Ruddy[d,e], Rhian Williams[d], Michael Perry[d], Matt Hitchings[a], Thomas S Wilkinson[a], Thomas Humphrey[a], M. Thomas P. Gilbert[b,f], Angharad Puw Davies[a,e#]

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

[b]Natural History Museum of Denmark, University of Copenhagen, Copenhagen K, Denmark

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

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

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

[f]Norwegian University of Science and Technology, University Museum, Trondheim, Norway.

#Corresponding Author: Angharad Davies:

Tel: 00 44 792 295034

e-mail: angharad.p.davies@swansea.ac.uk

Abstract

This study describes the analysis of DNA from heat-killed (boilate) isolates of *Mycobacterium tuberculosis* from two UK outbreaks where DNA was of sub-optimal quality for the standard methodologies routinely used in microbial genomics. An Illumina library construction method developed for sequencing ancient DNA was successfully used to obtain whole genome sequences, allowing analysis of the outbreak by gene-by-gene MLST, SNP mapping and phylogenetic analysis. All cases were spoligotyped to the same Haarlem H1 sub-lineage. This is the first described application of ancient DNA library construction protocols to allow whole genome sequencing of a clinical tuberculosis outbreak. Using this method it is possible to obtain epidemiologically meaningful data even when DNA is of insufficient quality for standard methods.

Keywords: *Mycobacterium tuberculosis*, whole genome sequencing, outbreak investigation, ancient DNA library construction, palaeogenomics.

Introduction

In 2017, tuberculosis incidence in Wales was 3.4 cases per 100 000 population [1], lower than the 2017 overall UK incidence of 8.4 cases per 100 000 population [2]. However, the structure of the National Health Service Trust responsible for health protection in Wales, Public Health Wales (PHW), links microbiology, public health and epidemiology into one organizational team. This, coupled with the relatively stable population, enables detailed analysis of links between cases and makes Wales an attractive place to study tuberculosis transmission dynamics.

During 2003-2005 an outbreak of *M. tuberculosis* occurred in a small town in south Wales (town G). Seven cases (GO1-GO7) were associated with a public house in the town. Case GO3, a barman in the public house, and case GO4, both had direct contact with all the other *M. tuberculosis* cases within the outbreak, at least six of whom had visited the public house. Case GO4 was considered to have been highly infectious, having been symptomatic for about 16 months before diagnosis. Both GO3 and GO4 were thought be the public health team to represent potential super-spreaders within the outbreak. During this period, the standard typing method in use by Public Health Wales was MIRU-VNTR. All the cases had identical MIRU-VNTR patterns. They were all fully susceptible to standard therapy.

In 2008, another outbreak was identified in an area of a nearby town, approximately 6km away (town T). There were close links between cases in the two outbreaks, with two of the cases in town T (TH1 and TH2) being direct contacts of GO3. TH2 was a regular at the public house at the centre of the outbreak in town G. However, MIRU-VNTR typing of the two groups of cases differed, with a polymorphism at a single MIRU-VNTR locus (MIRU16), suggesting the presence of two independent outbreaks within the area. Nonetheless the public health team felt it likely that all the cases formed part of one larger outbreak caused by the same strain of *M. tuberculosis* with divergence seen at locus MIRU16 due to a change in an endemic circulating strain. Epidemiologically linked isolates differing at fewer than two MIRU-VNTR loci have been suggested previously to be likely to be part of the same clonal complex within an outbreak [3] and MIRU-VNTR typing has been found to be unable to account for within-outbreak heterogeneity. It also provides limited information on the direction of transmission, identification of super-spreaders or outbreak origin [4, 5, 6, 7, 8, 9].

The development of affordable and accessible whole genome sequencing (WGS) protocols based around next generation sequencing (NGS) platforms such as the Illumina series, has

provided an alternative method for the investigation of *M. tuberculosis* outbreaks. The quality of the DNA sample is critical to the success of WGS. The *M. tuberculosis* samples for this study were provided by the Wales Centre for Mycobacteriology (WCM) in boilate form.

Although boilate extraction does release DNA, it is crude, inconsistent and yields DNA of lower integrity, in low quantity and of poorer quality in comparison with other extraction methods [10]. As might be expected, poor sample quality has a negative effect on the standard Nextera XT library preparation [11]. Given the need sometimes to generate WGS data from such samples, lessons might be learnt from the field of palaeogenomics, which attempts routinely to generate genome-scale data from nucleic acids that are highly fragmented, contaminated with non-target DNA, and often contain residual chemical impurities [12, 13, 14]. Recent developments in palaeogenomic methodologies have allowed WGS data to be obtained from a wide range of samples, spanning ancient humans and hominids [15, 16], mammals [17, 18], plants [19] and even pathogens [20] – many of which contain DNA with fragment lengths of <80bp. We therefore hypothesised that the application of palaeogenomic sequencing protocols might overcome the challenge of retrieving genomic data from the low purity and low-quality DNA of crude *M. tuberculosis* boilate samples.

Materials and Methods

Sample collection

Boilate samples from the *M. tuberculosis* isolates described above were obtained from the WCM, Public Health Wales, Llandough Hospital, Cardiff. The isolates had been cultured using the BACTEC™ MGIT™ 960 System (Becton Dickinson Diagnostic Systems, Sparks, MD) in containment level 3 facilities and then heat-killed by boiling for 35 minutes at 110^{o}C. MIRU-VNTR typing had been performed at the PHW Molecular Unit, University Hospital of Wales, Cardiff, with typing based on 15 loci, namely: ETRA, ETRB, ETRC, ETRD, ETRE, MIRU2, MIRU10, MIRU16, MIRU20, MIRU23, MIRU24, MIRU26, MIRU27, MIRU39 and MIRU40. Epidemiological information for each isolate was obtained through face-to-face interviews with a senior public health nurse from the original PHW outbreak investigation team, and from the outbreak documentation. All the cases in these outbreaks and under consideration here were fully susceptible to standard anti-tuberculosis chemotherapy.

Sequencing attempt using conventional Illumina protocols

Sequencing was first attempted directly from the boilates, following an ethanol precipitation. Indexed genomic DNA libraries were prepared for sequencing using the Illumina Nextera XT (V3) sample preparation protocol following the manufacturer's guidelines (2017 Illumina Inc., San Diego, CA, USA), size-selecting for fragments with an average size of 500bp. Bead-normalised sequencing libraries were pooled and sequenced on a MiSeq platform (2017 Illumina Inc., San Diego, CA, USA) using the V3 reagent kits and 600 cycles. The resulting paired-end reads were quality filtered with the Trimmomatic tool software [21] using a sliding window approach of 5 bases and a quality score of Q20 prior to contig assembly using the SPAdes genome assembler (Version 3.9.0) with K-mer sizes 33, 55, 77, 99 and 127 used [22]. In each case, this method failed to generate any sequence data at all, and unfortunately in this instance, replacement samples were not available. A method optimized for sequencing degraded DNA sources was therefore explored.

DNA extraction

The remaining *M. tuberculosis* boilates were transferred to tubes containing a 500µL solution of digestion buffer (10mM Tris-HCl pH8, 10mM NaCl , 5nM CaCl, 2.5mM EDTA, 1% SDS, 1% Proteinase K, and DTT) and 500µL of Phenol: Chloroform: Isoamyl alcohol solution (Sigma-Aldrich, St. Louis, MO, United States). Next, 0.6g of Zirconia/Silica beads (Cat. No. 11079105z, Biospec Products Inc., Bartlesville, OK, USA) were added to the tubes and each sample was homogenized using a TissueLyser II (Qiagen, Valencia, California), for 4 rounds

of 20 second bursts with cooling on ice for 30 seconds between rounds. After homogenization, samples were centrifuged for 10 minutes at 16,000 X g in a bench centrifuge to separate the phases. The aqueous upper phase (around 500µL) was gently transferred to a new low-bind 2mL Eppendorf tube and two volumes (1mL) of ice cold absolute ethanol (kept at -20C) were added to each sample. Samples were vortexed briefly and centrifuged again for 10 min at 16,000g. The supernatant was discarded and the tubes washed carefully with 700µL of 70% ethanol without disturbing the pellet. The ethanol wash was discarded and the pellet was left to dry for two minutes. Finally, the pellet was re-suspended and DNA eluted in buffer EB (Qiagen, Valencia, California), Extracted DNA was quantified on a Qubit fluorometer using a dsDNA high sensitivity assay (Life Technologies, Carlsbad, California) and an Agilent 2100 Bioanalyzer (Santa Clara, California). Following extraction, the samples were fragmented for 20 cycles in 30 second cycles within a Diagenode bioruptor 300.

Ancient DNA library preparation *protocol*

Carøe *et al.* [14] have recently published a new library construction protocol developed specifically for use on low concentration and degraded nucleic acid extracts. Based around a single tube blunt-end adaptor ligation, this so-called 'BEST' protocol has been shown to yield more complex libraries than other methods, due to removal of intermediate purification steps that generally lead to loss of the DNA molecules within the extract [14]. Illumina compatible libraries were constructed in this way at the laboratories of the Natural History Museum of Denmark, using 32 µl of extracted (see DNA extraction above) DNA from each boilate per sample as input. Based on qPCR results, libraries were indexed through PCR amplification for 10, 12 or 15 cycles, prior to visualisation and quantification on an Agilent Bioanalyser using the High Sensitivity DNA assay (Agilent technologies, Cheshire, UK). Subsequently, the indexed libraries were pooled at equimolar concentrations and then sequenced on an Illumina HiSeq 2500 platform (Illumina sequencing platforms, 2017) in 80bp single read mode by the Danish National High-Throughput DNA Sequencing Centre. The resulting single-end reads were quality filtered with the Trimmomatic tool [21] using a sliding window approach of 5 bases and a quality score of Q20 prior to contig assembly using the SPAdes genome assembler [22]. Raw reads for all the isolates are publically available (NCBI BioProject PRJNA556450).

Transmission chain

Cytoscape software [23] was used to generate a transmission tree.

Gene-by-gene MLST analysis

Gene-by-gene MLST analysis was carried out using Ridom SeqSphere Software [24]. A published core genome MLST (cgMLST) scheme [24, 6] was used for the analysis, which was based on 2891 core genes.

Whole Genome Sequence SNP mapping

Single nucleotide polymorphisms (SNPs) were identified using the standardised online CSI Phylogeny programme (Version 1.4; Call SNPs & Infer Phylogeny) of the Centre for Genomic Epidemiology (CGE) online tool [https://cge.cbs.dtu.dk/services/CSIPhylogeny/] [25]. A minimum spanning tree was constructed based on SNPs from 1123 sites from the WGS data using an adapted application of the Ridom SeqSphere software [24].

In silico *spoligotype*

Each isolate sequence was submitted to the Python-based SpolTyping [26] *in silico* software for prediction of spoligotype pattern. Resulting octal and binary patterns were then submitted to the SitVit database for determination of international typing assignment and assignment to globally recognised spoligtoype clades [27]. Additionally, spoligotype patterns were submitted to the TB-insight online server for identification of the corresponding *M. tuberculosis* lineage [27]. Isolates were assigned to major lineages based on the Conformal Bayesian Network (CBN) parameters, which employ a hierarchical Bayesian network based on PCR based biomarkers such as spoligotypes to classify isolates into given lineages [28].

Results

All the isolates were sequenced successfully using the aDNA sequencing protocol (Table 1). Sequence data for all isolates covered >99% of the core genes used in the cgMLST scheme. In addition, in all cases, sequence data covered >98% of the specified reference genome according to the CGE CSI phylogeny software. The minimum spanning tree (Figure 1) shows the genomic distances, based on allelic differences across 2891 core genes, between each of the isolates included in this dataset. A >12 allele difference was used as the threshold for exclusion from the outbreak [6].

cgMLST

Gene-by-gene cgMLST analysis appeared to indicate the presence of two outbreaks within the dataset, labelled outbreak 1 and outbreak 2, separated by 124 allelic differences. Outbreak 1 included the outbreak G public house cases GO2, GO3, GO4, GO5, GO6 and GO7 which all had fewer than 12 allelic differences between them. Within outbreak 1, two isolates could not be distinguished from one another, with no allelic differences detected (GO2 and GO6). A star-like topology is seen for outbreak 1, with GO2 and GO6 in the central position. Outbreak 2 is represented by three isolates, two from outbreak TH (TH1 and TH2) and from outbreak G, GO1 (Figure 1a). TH1 and GO1 could not be distinguished. Isolate TH2 diverged from GO1 and TH1 by only 4 allelic differences, well within the 12-allele limit and thus representing a direct transmission event [6]. Isolates GO8 and GO9, also from town G and so included in the figure for comparison, substantially exceeded the 12-allelic difference threshold for direct transmission with any other isolate within this dataset, consistent with the results of MIRU-VNTR typing, which previously found that GO8 and GO9 were not outbreak-related strains.

SNP mapping

SNP mapping also highlighted the presence of the same two outbreaks within this dataset (Figure 1b). For outbreak 1, a star-like topology was again present including isolates GO2, GO3, GO4, GO5, GO6 and GO7, as seen in the cgMLST analysis in Figure 1a. However, SNP mapping was able to distinguish GO2 and GO6 and found GO3 to be the central isolate. Outbreak 2 contained only two isolates, GO1 and TH1, with TH2 being excluded due to exceeding the threshold of 12 SNPs for outbreak inclusion. SNP mapping supported cgMLST and MIRU-VNTR in excluding GO8 and GO9 from either outbreak.

In silico *spoligotyping*

All the isolates were successfully predicted a spoligotype *in silico* (Table 2). The isolates could all be assigned to the same lineage and correlating spoligotype clades, the Euro-American lineage and Haarlem H1 spoligotype clades respectively. Three different spoligotypes and correlating international types were found: 47, 46 and 742. Isolates GO2 and GO4 had a different international type (742) from isolates GO3, GO5, GO6 and GO7 (international type 47) despite being directly linked within the same outbreak by both cgMLST and SNP mapping. In line with WGS, GO1 showed a closer association with the TH outbreak cases compared to the other town G cases, having the same spoligotype pattern (international type 46).

Discussion

This study describes the first application of an aDNA library construction protocol for the investigation of a clinical outbreak of *M. tuberculosis*. The 'BEST' ancient DNA library preparation protocol and subsequent sequencing were able to provide extensive sequence data from sub-optimal quality DNA from *M. tuberculosis* outbreak samples, where the standard protocol had failed. The aDNA library preparation protocol was able to circumvent the issue of short DNA fragment sizes and yield WGS data from *M. tuberculosi*s outbreak samples that would otherwise have been lost.

WGS provided extensive information on the outbreak. The application of gene-by-gene cgMLST provided similar, but not identical, results to those achieved by the SNP mapping procedure. Both agreed in assigning two separate outbreaks and both supported the epidemiological suspicion that a super-spreader was present, although the cgMLST and SNP mapping analyses conflicted in their assignment of the likely super-spreader. cgMLST indicated that the super-spreader was either GO2 or GO6, while traditional SNP mapping supported the assumption of the public health team that GO3 was the super-spreader responsible for multiple secondary cases within this outbreak. A possible reason for the discrepancy between the two methods could be the presence of gene families of a repetitive nature being included in the analyses, such as those for PE_PPE which show disproportionately high amount of divergence [29]. cgMLST removed these regions and this could at least partly explain the discrepancies. It is important to note that most SNP-calling procedures in *M. tuberculosis* epidemiology filter out repetitive regions, and that therefore the procedure of mapping and SNP-calling without filtering, described here, is not directly comparable with that described in other studies.

In silico application of SpoTyping software was achieved without further laboratory work and at no extra cost to initial sequencing data [26]. Each isolate in this outbreak had a spoligotype pattern corresponding to the Haarlem H1 clade. It has been demonstrated previously that the value of inferring a recent common ancestor of isolates within potentially related outbreaks, with identification of a causative circulating strain being a common feature [8, 30]. Previous studies of *M. tuberculosis* in the Inuit and Greenland populations of North America, which have stable populations, have also documented the cause of multiple outbreaks within the region as the ongoing spread of an evolving founder strain that has continually spread across the area over decades [8, 30].

The results of *in silico* spoligotyping, together with the strong epidemiological links between the two outbreaks, lends support to the public health team's case that both G and T outbreak cases were part of the same on-going outbreak. The apparent existence of two outbreaks may be due to the absence of intermediate isolates not included in this dataset coupled with recent minor genomic changes [8], a conclusion consistent with the presence of a polymorphism at only one MIRU-VNTR locus between the two genotypes. Such problems with MIRU-VNTR typing have been described before [4, 8, 30]. Approximately one-third of reported tuberculosis cases are culture-negative, so that there is no isolate for analysis.

The work described here demonstrates that when necessary, clinically useful data can be obtained from sub-optimal quality samples by applying an ancient DNA library construction protocol to overcome the need for DNA of high purity and quality. The success of the 'BEST' aDNA library construction protocol here highlights a clinical application for a method previously associated only with palaeogenomic studies, and shows how the transfer of such techniques in defined circumstances could provide clinical benefit.

Acknowledgments

This work was funded by St. David's Medical Foundation & Coleg Cenedlaethol Cymraeg funding to APD & RJ, a EMBO Short-term Fellowship to RJ, and an ERC Consolidator Grant (681396- Extinction Genomics) to MTPG. The authors would like to thank Christian Carøe and other members of the Gilbert laboratory at the Natural History Museum, Denmark for technical guidance.

271

272

References

1. Public Health Wales Communicable Disease Surveillance Centre. 2018. Tuberculosis in Wales Annual Report.
2. Public Health England 2018. Reports of cases of tuberculosis to enhanced tuberculosis surveillance systems: United Kingdom, 2000 to 2017.
3. Allix-Béguec C, Harmsen D, Weniger T, Supply P, Niemann S. 2008. Evaluation and strategy for use of MIRU-VNTRplus, a multifunctional database for online analysis of genotyping data and phylogenetic identification of *Mycobacterium tuberculosis* complex isolates. J Clin Microbiol. 46:2692-2699.
4. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dedicoat M, 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 Inf Dis. 13:137-146.
5. Walker TM, Kohl TA, Omar SV, Hedge J, Elias CDO, Bradley P, Iqbal Z, Feuerriegel S, Niehaus KE, Wilson DJ. 2015. Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. Lancet Inf Dis.15:1193-1202.
6. Kohl TA, Diel R, Harmsen D, Rothgänger 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.
7. Takiff HE & Feo O. 2015. Clinical value of whole-genome sequencing of Mycobacterium tuberculosis. Lancet Inf Dis. 15:1077-1090.
8. Bjorn-Mortensen K, Soborg B, Koch A, Ladefoged K, Merker M, Lillebaek T, Andersen A, Niemann S, Kohl T. 2016. Tracing *Mycobacterium tuberculosis* transmission by whole genome sequencing in a high incidence setting: a retrospective population-based study in East Greenland. Sci Rep. 6:33180.
9. Gardy JL, Johnston JC, Sui SJH, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma M, Elwood K, Jones S, Brinkman F, Brunham R, Tang P. 2011. Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. NEJM. 364:730-739.
10. Aldous WK, Pounder JI, Cloud JL, Woods GL. 2005. Comparison of six methods of extracting *Mycobacterium tuberculosis* DNA from processed sputum for testing by quantitative real-time PCR. J Clin Microbiol. 43:2471-2473.

11. Tyler AD, Christianson S, Knox NC, Mabon P, Wolfe J, Van Domselaar G, Graham MR, Sharma MK. 2016. Comparison of sample preparation methods used for the next-generation sequencing of *Mycobacterium tuberculosis*. PloS ONE. 11:e0148676.
12. Schubert M, Ermini L, Der Sarkissian C, Jónsson H, Ginolhac A, Schaefer R, Martin MD, Fernandez R, Kircher M, McCue M, Willersely E, Orlando L. 2014. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nat Protoc. 9:1056-82.
13. Orlando L, Gilbert MT, Willerslev E. 2015. Reconstructing ancient genomes and epigenomes. Nat Rev Genet. 16:395-408.
14. Carøe CGS, Gopalakrishnan S, Vinner L, Mak SS, Sinding MH, Samaniego JA, Wales N, Sicheritz-Ponten T, Gilbert MTP. 2017. Single-tube library preparation for degraded DNA. Methods Ecol Evol. doi10.1111/2041-210X.12871
15. Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, Gupta R, Bertalan M, Nielsen K, Gilbert MT, Wang Y, Raghavan M, Campos PF, Kamp HM, Wilson AS, Gledhill A, Tridico S, Bunce M, Lorenzen ED, Binladen J, Guo X, Zhao J, Zhang X, Zhang H, Li Z, Chen M, Orlando L, Kristiansen K, Bak M, Tommerup N, Bendixen C, Pierre TL, Grønnow B, Meldgaard M, Andreasen C, Fedorova SA, Osipova LP, Higham TF, Ramsey CB, Hansen TV, Nielsen FC, Crawford MH, Brunak S, Sicheritz-Pontén T, Villems R, Nielsen R, Krogh A, Wang J, Willerslev E. 2010. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature. 463:757-62.
16. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, Li H, Mallick S, Dannemann M, Fu Q, Kircher M, Kuhlwilm M, Lachmann M, Meyer M, Ongyerth M, Siebauer M, Theunert C, Tandon A, Moorjani P, Pickrell J, Mullikin JC, Vohr SH, Green RE, Hellmann I, Johnson PL, Blanche H, Cann H, Kitzman JO, Shendure J, Eichler EE, Lein ES, Bakken TE, Golovanova LV, Doronichev VB, Shunkov MV, Derevianko AP, Viola B, Slatkin M, Reich D, Kelso J, Pääbo S. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 505,43-9.
17. Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, Schubert M, Cappellini E, Petersen B, Moltke I, Johnson PL, Fumagalli M, Vilstrup JT, Raghavan M, Korneliussen T, Malaspinas AS, Vogt J, Szklarczyk D, Kelstrup CD, Vinther J, Dolocan A, Stenderup J, Velazquez AM, Cahill J, Rasmussen M, Wang X, Min J,

Zazula GD, Seguin-Orlando A, Mortensen C, Magnussen K, Thompson JF, Weinstock J, Gregersen K, Røed KH, Eisenmann V, Rubin CJ, Miller DC, Antczak DF, Bertelsen MF, Brunak S, Al-Rasheid KA, Ryder O, Andersson L, Mundy J, Krogh A, Gilbert MT, Kjær K, Sicheritz-Ponten T, Jensen LJ, Olsen JV, Hofreiter M, Nielsen R, Shapiro B, Wang J, Willerslev E. 2013. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature. 499:74-8.

18. Miller W, Drautz DI, Ratan A, Pusey B, Qi J, Lesk AM, Tomsho LP, Packard MD, Zhao F, Sher A, Tikhonov A, Raney B, Patterson N, Lindblad-Toh K, Lander ES, Knight JR, Irzyk GP, Fredrikson KM, Harkins TT, Sheridan S, Pringle T, Schuster SC. 2008. Sequencing the nuclear genome of the extinct woolly mammoth. Nature. 456:387-90.
19. Mascher M., Schuenemann VJ, Davidovich U, Marom N, Himmelbach A, Hübner S, Korol A, David M, Reiter E, Riehl S, Schreiber M, Vohr SH, Green RE, Dawson IK, Russell J, Kilian B, Muehlbauer GJ, Waugh R, Fahima T, Krause J, Weiss E, Stein N. 2016. Genomic analysis of 6,000-year-old cultivated grain illuminates the domestication history of barley. Nat Genet. *48:*1089-93.
20. Schuenemann VJ, Singh P, Mendum TA, Krause-Kyora B, Jäger G, Bos KI, Herbig A, Economou C, Benjak A, Busso P, Nebel A, Boldsen JL, Kjellström A, Wu H, Stewart GR, Taylor GM, Bauer P, Lee OY, Wu HH, Minnikin DE, Besra GS, Tucker K, Roffey S, Sow SO, Cole ST, Nieselt K, Krause J. 2013. Genome-wide comparison of medieval and modern Mycobacterium leprae. Science. 341:179-83.
21. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30:2114-2120
22. 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.
23. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13:2498-2504.
24. 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.

25. Kaas RS, Leekitcharoenphon P, Aarestrup FM, Lund, O. 2014. Solving the problem of comparing whole bacterial genomes across different sequencing platforms. PLoS ONE. 9:e104984.
26. Xia E, Teo YY, Ong RTH. 2016. SpoTyping: fast and accurate *in-silico Mycobacterium* spoligotyping from sequence reads. Genome Med. 8:19.
27. Shabbeer A, Cowan LS, Ozcaglar C, Rastogi N, Vandenberg SL, Yener B, Bennett KP. 2012. TB-Lineage: an online tool for classification and analysis of strains of *Mycobacterium tuberculosis* complex. Inf Genet Evol. 12:789-797.
28. Aminian M, Shabbeer A, Bennett K. 2010. A conformal Bayesian network for classification of *Mycobacterium tuberculosis* complex lineages. BMC Bioinformatics. 11:S4
29. Roetzer A, Diel R, Kohl TA, Ruckert C, Nubel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rusch-Gerdes S, Supply P, Kalinowski J, Niemann S. 2013. Whole Genome Sequencing versus Traditional Genotyping for Investigation of a *Mycobacterium tuberculosis* Outbreak: A Longitudinal Molecular Epidemiological Study. PLOS Med. 10: e1001387
30. Lee RS, Radomski N, Proulx JF, Levade I, Shapiro BJ, McIntosh F, Soualhine H, Menzies D, Behr MA. 2015. Population genomics of *Mycobacterium tuberculosis* in the Inuit. Proc Nat Acad Sci. 112:13609-13614.

Table 1: Ancient DNA protocol sequencing of nine outbreak isolates. The percentage of the core genome MLST genes present is shown as well as the percentage of the reference *M. tuberculosis* H37Rv genome covered according to the CSI phylogeny algorithm provided by the Centre for Genomic Epidemiology [7].

Sample ID	No. contigs	Largest contig	Total length	N50	Average depth of coverage	% cgMLST	% reference genome covered	GenBank Accession	% Error rate
TH1	151	174895	4372963	64226	177.0	99.52	98.27	VOGE00000000	0.45
TH2	181	171547	4347932	61791	86.9	99.34	98.18	VOGF00000000	0.46
GO1	150	174836	4360310	69175	159.0	99.52	98.45	VOGD00000000	0.44
GO2	163	174750	4353499	63534	90.1	99.52	98.07	VOGG00000000	0.44
GO3	162	211047	4401955	75381	221.2	99.52	98.26	VOGH00000000	0.45
GO4	143	257745	4357032	69538	129.0	99.41	98.15	VOGI00000000	0.42
GO5	134	228152	4363387	76515	158.8	99.45	98.50	VOGJ00000000	0.43
GO6	148	210603	4362517	72538	125.0	99.48	98.36	VOGK00000000	0.40
GO7	161	174721	4349894	64061	76.3	99.52	98.50	VOGL00000000	0.41

Table 2: *in silico* spoligotyping results for each of the outbreak isolates. Results include the predicted spoligotype (produced by SpolDB4), international spoligotype (SITVIT database), lineage assignment and clade assignment (both outputted from the TB-insight online server).

Isolate	Predicted Spoligotype	International type	Lineage	Clade
GO3	777777774020771	47	Euro-American	H1
TH1	777777770000000	46	Euro-American	H1
TH2	777777770000000	46	Euro-American	H1
GO1	777777770000000	46	Euro-American	H1
GO2	777777770020771	742	Euro-American	H1
GO4	777777770020771	742	Euro-American	H1
GO5	777777774020771	47	Euro-American	H1
GO6	777777774020771	47	Euro-American	H1
GO7	777777774020771	47	Euro-American	H1

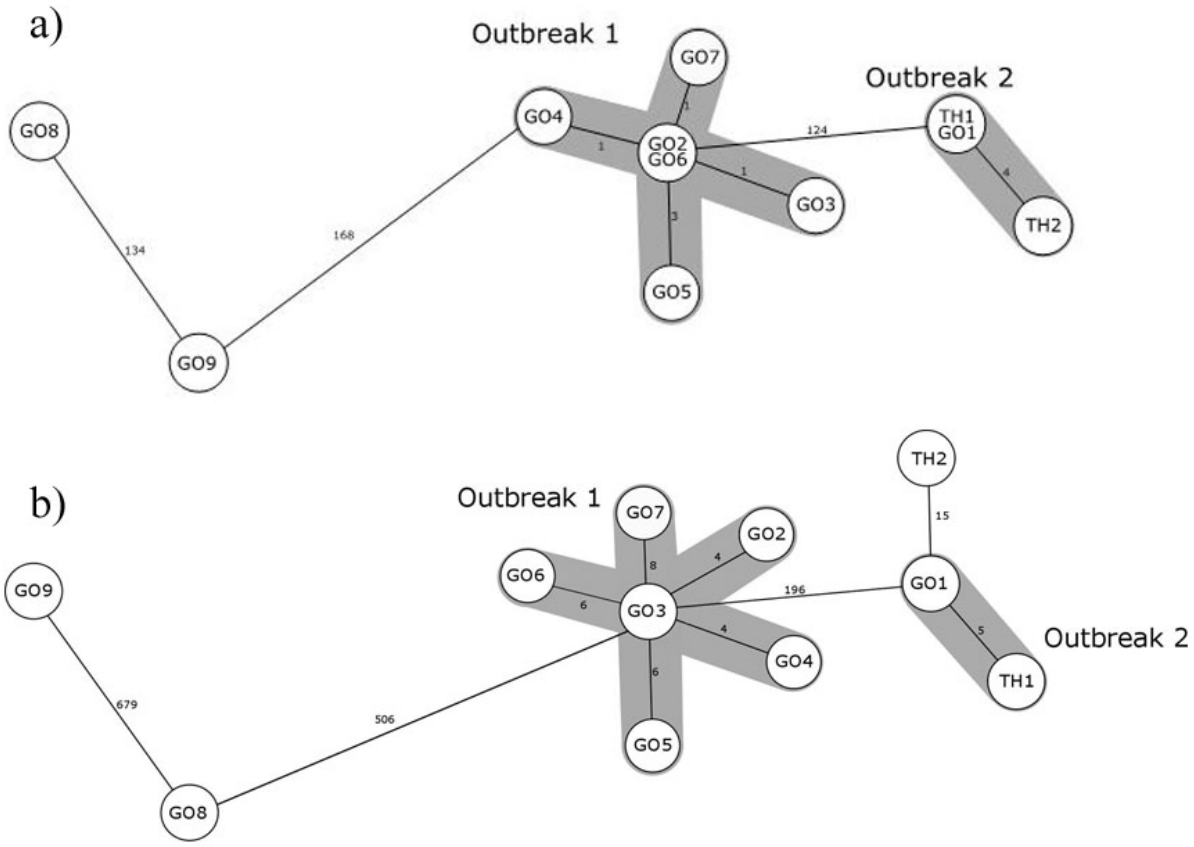
406 **Figure Legend**

407 **Figure 1:** Results from analysing the nine isolates, with the addition of 2 isolates (GO8 &
408 GO9) from the same geographical area isolated during the same time-period, previously
409 sequenced by a standard method. **a**: cgMLST minimum spanning tree of the 11 isolates
410 with numbers representing the number of allelic differences between isolates; and **b**: CSI
411 phylogeny based minimum spanning tree of the 11 isolates on a total of 1123 SNPs, with
412 numbers representing the number of SNPs between isolates. Branches between isolates
413 are not to scale. Shaded areas represent those associated with the labelled outbreak.
414 Isolates with prefix TH were from Town T; prefix GO indicates isolates from Town G.

415

416

Figure 1


a)
Outbreak 1
Outbreak 2
GO8
GO4
GO7
GO2
GO6
TH1
GO1
TH2
GO3
GO5
GO9
134
168
1
1
124
1
4
3
b)
Outbreak 1
Outbreak 2
TH2
GO7
GO2
GO9
GO6
GO3
GO1
GO4
TH1
GO5
GO8
15
8
4
196
6
5
4
6
679
506