A bioinformatic analysis of *Mycobacterium tuberculosis* and host genomic data

Jody Emile Phelan

Thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy

University of London

September 2017

Department of Pathogen Molecular Biology

Faculty of Infectious Tropical Disease

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

Funded by the BBSRC

Research group affiliation(s): Taane G Clark, Martin L Hibberd & Sanjib Bhakta

I, Jody Emile Phelan, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in this thesis.

Signed _________________ Date _________________

Abstract

Human tuberculosis disease (TB) is caused by bacteria within the *Mycobacterium tuberculosis* complex, including *M. tuberculosis* (*Mtb*). Genetic variation within the pathogen can lead to drug resistance, affect virulence and transmissibility. I have analysed *Mtb* whole genome sequence data to improve the understanding of global genetic variation, and the resulting insights could ultimately assist the development of TB control measures.

Whole genome sequencing platforms are being used to infer drug resistance profiles, and thereby could assist clinical management. I investigated the reproducibility of sequence data from two platforms (Illumina MiSeq, Ion Torrent PGM™) and two rapid analytic pipelines (TBProfiler, Mykrobe predictor). DNA replicates from the reference strain (H37Rv) and 10 drug-resistant strains were sequenced, and inferred drug resistance genotypes were compared to drug susceptibility testing phenotypes.

Genome-wide association study (GWAS) can be used to detect mutations associated with *Mtb* drug resistance. A first GWAS (n=127) attempted to identify mutations associated with minimum inhibitory concentrations for first-line anti-tuberculosis drugs. A second GWAS was applied to a large global set (n>6400) to identify mutations associated with first- and second-line drug resistance.

M. aurum is an environmental mycobacteria that has been proposed as a model for the development of anti-TB drugs. I have assembled and annotated its draft genome, and identified copy number variants in known drug resistance targets.

Approximately 10% of the *Mtb* genome consists of two gene families (*pe/ppe*) that are poorly characterised, and are hypothesised to be important virulence factors. Using a *de novo* assembly approach, I characterised these genes and their diversity across a global collection of clinical isolates with high depth short-read sequence data (n=518). A follow-up study using a long-read sequence technology (n=18, diverse stain types) confirmed the findings. This work also generated new annotated reference genomes and characterised methylation sites, which may affect transmissibility, pathogenicity and virulence.

A future direction of the TB genomics field is to identify genetic check points in host-pathogen interactions using both human and *Mtb* genotypes. I analysed the genomes of ~720 TB case–*Mtb* pairs and identified susceptibility markers, which are promising targets for future control measures.

Acknowledgements

I would like to thank my supervisors Taane Clark, Martin Hibberd and Sanjib Bhakta.

The BBSRC and members of the LIDO team.

Members of the hub - Ernest, Matt, Yaa and Ben.

Colleagues who have helped me during my PhD - Susana Campino, Francesc Coll, Stephane Hue, Nick Furnham, David Ascher, Douglas Pires, Stephanie Portelli.

Sammy Assefa for helping me both at LSHTM and at Barclays.

Nuno Sepulveda for all the advice and good Birkbeck times.

I thank my family for all their support they have given me on the road to get where I am today - My mother and father, Siobhan, Jessica and Lisette.

Akash Maniam - they guy who has been there since the start of my London Journey.

Members of Hendon Massive - Amit Melwani, Asad Rahman and Jayson Parmar.

Archie Khan for always being there for me throughout my PhD.

Additional Publications

I contributed to other manuscripts which were not part of my PhD:

*Andreu, N., **Phelan, J.**, de Sessions, P.F., Cliff, J.M., Clark, T.G., Hibberd, M.L., 2017. Primary macrophages and J774 cells respond differently to infection with Mycobacterium tuberculosis. Sci. Rep. 7, 42225. doi:10.1038/srep42225*

*de Oliveira, T.C., Rodrigues, P.T., Menezes, M.J., Gonçalves-Lopes, R.M., Bastos, M.S., Lima, N.F., Barbosa, S., Gerber, A.L., Loss de Morais, G., Berná, L., **Phelan, J.**, Robello, C., de Vasconcelos, A.T.R., Alves, J.M.P., Ferreira, M.U., 2017. Genome-wide diversity and differentiation in New World populations of the human malaria parasite Plasmodium vivax. PLoS Negl. Trop. Dis. 11, e0005824. doi:10.1371/journal.pntd.0005824*

*Dheda, K., Limberis, J.D., Pietersen, E., **Phelan, J.**, Esmail, A., Lesosky, M., Fennelly, K.P., te Riele, J., Mastrapa, B., Streicher, E.M., Dolby, T., Abdallah, A.M., Ben-Rached, F., Simpson, J., Smith, L., Gumbo, T., van Helden, P., Sirgel, F.A., McNerney, R., Theron, G., Pain, A., Clark, T.G., Warren, R.M., 2017. Outcomes, infectiousness, and transmission dynamics of patients with extensively drug-resistant tuberculosis and home-discharged patients with programmatically incurable tuberculosis: a prospective cohort study. Lancet Respir. Med. 5, 269–281. doi:10.1016/S2213-2600(16)30433-7*

*Kanji, A., Hasan, R., Ali, A., Zaver, A., Zhang, Y., Imtiaz, K., Shi, W., Clark, T.G., McNerney, R., **Phelan, J.**, Rao, S., Shafiq, S., Hasan, Z., 2017. Single nucleotide polymorphisms in efflux pumps genes in extensively drug resistant Mycobacterium tuberculosis isolates from Pakistan. Tuberculosis 107, 20–30. doi:10.1016/j.tube.2017.07.012*

*Refrégier, G., Abadia, E., Matsumoto, T., Ano, H., Takashima, T., Tsuyuguchi, I., Aktas, E., Cömert, F., Gomgnimbou, M.K., Panaiotov, S., **Phelan, J.**, Coll, F., McNerney, R., Pain, A., Clark, T.G., Sola, C., 2016. Turkish and Japanese Mycobacterium tuberculosis sublineages share a remote common ancestor. Infect. Genet. Evol. 45, 461–473. doi:10.1016/j.meegid.2016.10.009*

*Sengstake, S., Bergval, I.L., Schuitema, A.R., de Beer, J.L., **Phelan, J.**, de Zwaan, R., Clark, T.G., van Soolingen, D., Anthony, R.M., 2017. Pyrazinamide resistance-conferring mutations in pncA and the transmission of multidrug resistant TB in Georgia. BMC Infect. Dis. 17, 491. doi:10.1186/s12879-017-2594-3*

*Campos, M. C., **Phelan, J**., Francisco, A. F., Taylor, M. C., Lewis, M. D., Pain, A., Clark, T. G. & Kelly, J. M. Genome-wide mutagenesis and multi-drug resistance in American trypanosomes induced by the front-line drug benznidazole. Sci. Rep. **7,** 14407 (2017).*

Table of Contents

Abbreviation list

AMK	Amikacin
BCG	Bacillus Calmette-Guérin
CAP	Capreomycin
CIP	Ciprofloxacin
CYS	Cycloserine
DOTS	Directly observed treatment short course
DR	Direct repeat
DST	Drug susceptibility testing
EAI	East-African-Indian
EMB	Ethambutol
ENA	European nucleotide archive
ETH	Ethionamide
FLQ	Fluoroquinolone
GWAS	Genome wide association studies
HBV	Hepatitis B virus
HIV	Human immune deficiency virus
IFNγ	interferon-gamma
indel	Insertions and deletions
INH	Isoniazid
IPD	Inter-pulse duration
KAN	Kanamycin
LAM	Latin Americal Mediterranean
LOF	Loss of function
LSPs	long sequence polymorphisms
MDR	Multidrug-resistant
MGIT	Mycobacterial Growth Indicator Tube
MHC	Major histocompatibility complex
MIC	Minimum inhibitory concentration
MIRU	Mycobacterial interspersed repetitive units
MODS	Microscopic Observation Drug Susceptibility assay
MOX	Moxifloxacin
MTase	Methyl transferase
Mtb	*Mycobacterium tuberculosis*
MTBC	*Mycobacterium tuberculosis* complex
NGS	Next Generation Sequencing
NK-cells	Natural killer-cells
OFL	Ofloxacin
PacBio	Pacific Biosciences
PAS	Para-aminosalicylic acid
PCA	Principal component analysis
PDB	Protein data bank

PDIM	Phthiocerol dimycocerosate
PE	Proline-glutamate
PPE	Proline-proline-glutamate
PZA	Pyrazinamide
RDs	Regions of difference
RIF	Rifampicin
RRDR	Rifampicin resistance-determining region
SLID	Second line injectable drugs
SNP	Single nucleotide polymorphism
STM	Streptomycin
T-cells	T Lymphocytes
TB	Tuberculosis disease
VCF	Variant call file
VNTR	Variable number of tandem repeats
WHO	World Health Organisation
XDR	Extensively drug-resistant

Chapter 1

Introduction

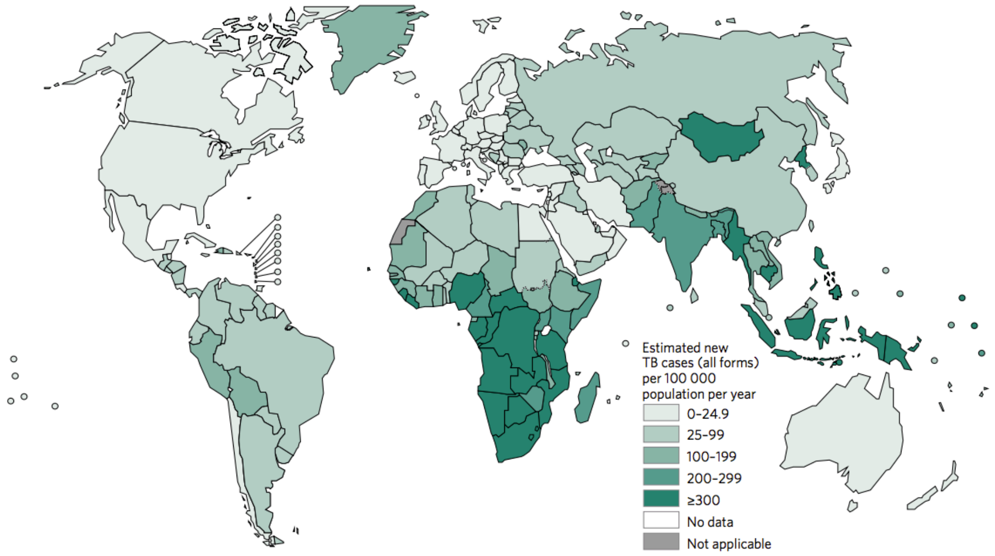
1. Introduction

1.1 *Global burden of tuberculosis disease*

Human tuberculosis disease (TB) is a major global public health problem, with an estimated 10.4 million new cases and 1.8 million deaths in 2015 alone[1]. Disease control is becoming difficult due to increasing drug resistance and in some populations HIV co-infection[1]. The majority of new cases (60%) were from China, India, Indonesia, Nigeria, Pakistan and South Africa[1] (**Figure 1**). HIV infection increases the incidence and mortality risk of the disease, and 1.2 million (11%) of new cases were HIV-positive[1]. One of the aims set out by the End TB Strategy is a 35% reduction of TB deaths by 2020 compared to 2015. To achieve this, a reduction of 4-5% needs to be sustained annually[1].

Figure 1

The estimated incidence of new TB cases per annum (per 100,000). Taken from the WHO Global Tuberculosis Report[1].

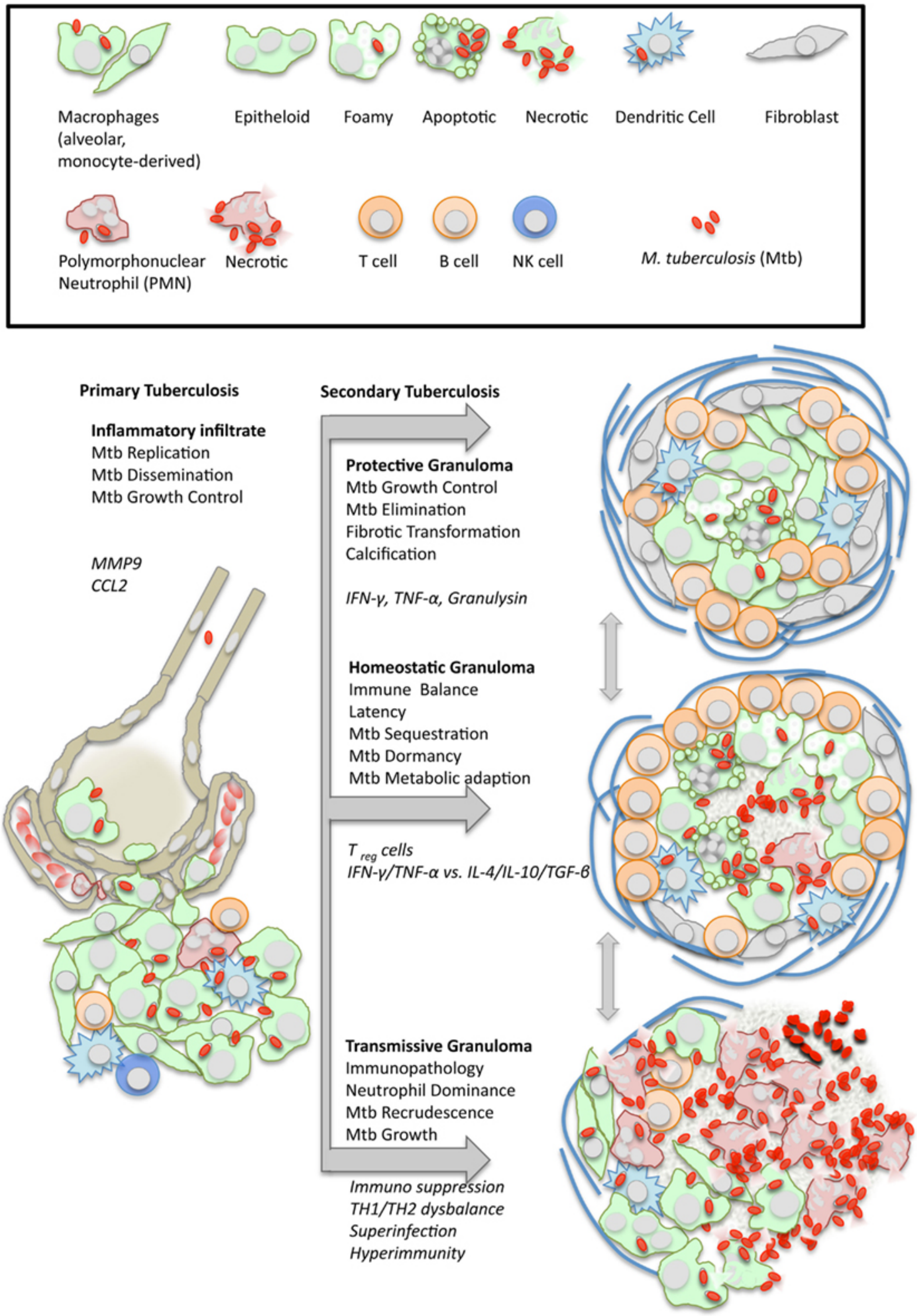

Estimated new TB cases (all forms) per 100 000 population per year
0–24.9
25–99
100–199
200–299
≥300
No data
Not applicable

1.2 *Disease etiology, risk factors and host susceptibility*

Symptoms of TB include weight loss, chest pain, coughing, fever and night sweats. Droplets containing the bacterium are inhaled, reach the alveoli and invoke an immune response though macrophages and granulocytes. Once the bacilli have entered the lung they are engulfed by macrophages which move into deeper tissue. *Mtb* then replicates within the macrophage – eventually causing apoptosis and rupture of the host cell. More macrophages are recruited to engulf the debris and the granuloma is formed through the recruitment of NK-cells and T-cells (**Figure 2**)[2]. This process can either lead to replication of the bacilli in macrophages and progression to active disease or the bacilli are contained and replication is limited, resulting in latent infection. Pulmonary TB is the most common form of TB due to the lungs being the primary point of infection but the bacilli can spread to other parts of the body leading to extra pulmonary TB. Only 10% of individuals infected will develop active TB. In addition to pathogen factors such as lineage and drug resistance, there are many host risk factors involved with the outcome of infection including age, HIV status, immunosuppression and genetics. There is some evidence of inter-population variation of resistance levels to TB[3]. This variation could be attributed to differing socio-economic levels of populations, but also host genetic factors affecting ethnic group susceptibility[4]. Several twin studies have suggested that susceptibility may be heritable, with higher concordance for the development of TB among mono- as opposed to di-zygotic twins[5]. Using experimental and molecular approaches, some genetic differences between populations contributing towards the altered level of infection have been found. Altered interferon-gamma (*IFNγ*) expression in response to mycobacterial antigens has been implicated in infection with atypical mycobacteria[7]. Studies of patients suffering from high susceptibility to mycobacteria led to a number of variants in

Figure 2

The initial stages of infection up to the formation of the granuloma. Adapted from Ehlers *et al*[6]. Initial infection can lead to i) clearance, ii) latent infection or iii) active TB. This figure demonstrates the many different cell types are involved in the generation of the granuloma.



receptors and ligands of the IL-12/IFNγ pathway being implicated in susceptibility[8–12]. These studies, while useful in demonstrating that host genetics is a factor in susceptibility to TB, analysed rare variants which would not be expected to exist at high frequencies in human populations. Identification of other host genetic variants affecting susceptibility is vital, especially to improve patient management, and genome wide association studies (GWAS) using large numbers of polymorphisms have been employed. The GWAS approach has worked well for other infectious diseases, with associated loci detected for Leprosy, Dengue, HIV and HBV[13]. The first GWAS was performed using 11,425 individuals in a combined cohort from Ghana and The Gambia[12]. They found a region on 18q11.2 to be highly significant with susceptibility to TB. However, validation studies in other cohorts have reported conflicting results[10,14]. The largest study to date analysed 5,530 TB cases and 5,607 healthy controls in a Russian population and identified 7 SNPs on chromosome 8 in the *ASAP1* gene as being associated with susceptibility to TB [9]. However a subsequent study in a Chinese Han population could not replicate the strong associations reported[8]. This inability to replicate results has become a consistent theme for GWASs in the TB field. To date there have not been any candidate regions in the human genome which have been significantly associated with TB susceptibility across the majority of genome wide association studies. Human or pathogen (or both) variation across regions may be an important component influencing the chances of genetic association reproducibility.

1.3 *Diagnosis*

Active and latent TB have different diagnosis methods. The recommended diagnosis method for latent TB is the tuberculin skin test or the blood interferon-gamma release assay. The former is more cost effective and may be more suited for low-income regions while the latter has an improved specificity[15]. Both these tests measure the response of the adaptive immune system and therefore need to be administered greater than eight weeks' post infection to produce a reliable result. Symptoms and signs of TB like lesions in the lung can checked using chest x-ray and used to diagnose TB. However, symptoms can overlap with other more common diseases in low endemic countries and do not immediately point towards TB. Additionally, interpretation of x-rays requires experience and is subjective. Smear microscopy is the most widely used method of tuberculosis diagnosis for active TB. Its sensitivity is estimated at ~70% and can drop to as low as ~35% in some clinical settings[16,17]. Culture is the gold standard for diagnosis of tuberculosis. This can be performed using n solid (> 3-4 weeks) or liquid culture (10-14 days)[18]. Molecular diagnostic tests on sputum samples, such as the Xpert MTB/RIF assay can detect *Mtb* with an increased sensitivity and can also detect resistance to certain drugs, in this case rifampicin[15]. Recent efforts have concentrated on detecting *Mtb* nucleic acids by sequencing direct from sputum[19] or with limited culturing[20]. Informatics tools have been developed to rapidly profile the *Mtb* for drug resistance and strain-type from sequence data, thus potentially improving patient management[21].

1.4 *Treatments*

Although there is no perfect treatment for TB, there are many anti-tuberculous drugs, which are subdivided into two main categories. The first line drugs include rifampicin, isoniazid, ethambutol, streptomycin and pyrazinamide. Second line drugs such as fluoroquinolones and injectables should be supplemented when treatment fails with the above (**Table 1**). Other new drugs currently used in clinical trials include bedaquiline, delamanid and clofazimine, which are important for the most serious drug resistance cases. The WHO, and the UK based National Institute for health and Care Excellence, standard recommended regimen for active respiratory TB consists of a 6 month regimen of the first line drugs isoniazid and rifampicin supplemented with ethambutol and pyrazinamide for the first two months[22] and can achieve positive outcome rates of up to 95%[23]. Latent infections can be treated as a preventative measure in high prevalence areas. A regimen of isoniazid for 6-9 months is recommended[15].

Table 1 Summary of the drugs used to treat various forms of tuberculosis. Adapted from Zumla *et al*[15]

Drugs	Drug regimen
Standard regimen	6 months rifampicin and isoniazid supplemented by ethambutol and pyrazinamide in the first two months
Latent infection	6-9 months isoniazid
Multi- drug resistant TB	Addition of second-line drugs e.g.: • Fluoroquinolones – ofloxacin • Injectables – kanamycin • Bacteriostatic – para-aminosalicylic acid
Extensively drug resistant TB, use of new drugs	Bedaquiline, delamanid and clofazimine

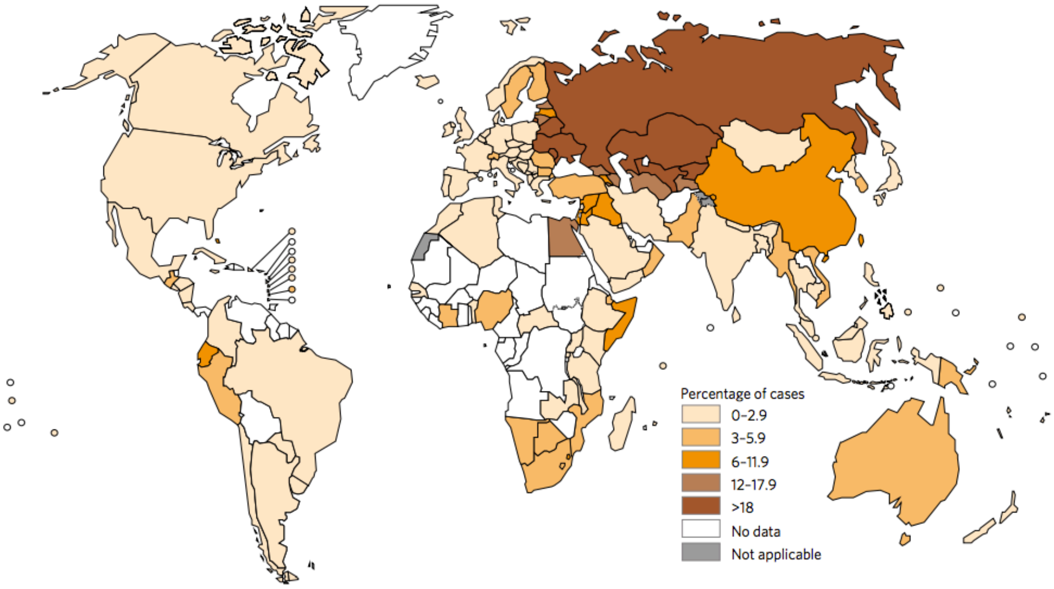
When treatment failure occurs with first line drugs, second line regimens must be used. Multi-drug resistant tuberculosis (MDR-TB) is defined by resistance to at least rifampicin and isoniazid. Additional resistance to second-line drugs, the fluoroquinolones and injectables is denoted as extensively drug resistance (XDR-TB). M/XDR-TB regimens are complicated and costly. In particular, the current World Health Organisation (WHO) approved MDR-TB regimen has an overall efficacy of only 50% and the median time to culture conversion can be as long as five months[15]. The current regimen is also toxic and six months of painful injections promotes non-adherence. Drug susceptibility testing (DST) must be used to inform on which first/second line drugs to remove and supplement. The toxicity of current second line drugs and the development of resistance has created a need for new drugs. Several new drugs are currently being evaluated for use. Bedaquiline and delamanid are two new drugs which have undergone phase II and III trials[24–26] and have been assigned the been assigned as "add-on agents", to be used in complicated cased of MDR-TB[27]. These drugs, along with several more in preclinical and clinical trial stages[28] will aid in the fight against MDR-TB and XDR-TB.

1.5 *Drug resistance*

Mtb drug resistance is conferred by the accumulation of mutations (single nucleotide polymorphisms (SNPs), insertions and deletions (indels)) in genes coding for drug-targets or -converting enzymes[29,30]. To overcome a loss of fitness that arises during the accumulation of such mutations[31], putative compensatory mechanisms have been described[32,33]. Ineffective use of the drugs, such as defaulting from treatment, can cause the host *Mtb* population to go through a partial population bottleneck and lead to the mutants to increase in frequency - effectively causing all *Mtb* within a patient to become resistant. Mutations conferring resistance to rifampicin and isoniazid are well characterised. Rifampicin binds to and inhibits RNA polymerase[34]. Resistance to rifampicin is caused by mutations in the *rpoB* gene coding for RNA polymerase β subunit[35]. Nearly all resistance conferring mutations occur in an 81bp region in *rpoB* called the rifampicin resistance-determining region (RRDR)[35]. Isoniazid is a prodrug which is activated by KatG catalase-peroxidase and binds to InhA to inhibit mycolic acid synthesis[36]. Mutations in the *katG* gene or in the *inhA* promoter lead to resistance[37,38]. Mutations for second line drugs are less well characterised and as a result are difficult to predict using sequencing[21]. MDR-TB and XDR-TB cases have been reported in 117 countries[1]. Approximately 3.9% of new cases and 21.0% of previously treated cases were estimated to have MDR-TB. Additionally, an estimated 9.5% of MDR-TB cases are XDR-TB[1].

Figure 3

A map indicating the percentage of new cases with MDR-TB. High incidence is seen in Russia and other ex-Soviet Republics. Taken from the 2016 WHO Global Tuberculosis Report[1].



The burden of MDR-TB is especially high in India, China, South Africa and Russia (**Figure 3**). It was estimated that of only 20.0% new MDR-TB cases eligible for treatment were enrolled in appropriate programs[1]. One of the five priority actions of the WHO to address MDR-TB is through "expansion of rapid testing and diagnosis of MDR-TB cases"[23]. As mentioned above, some molecular diagnostic tests can confirm some of the drug resistance genetic markers, including the use of the Xpert MTB/RIF assay. However, standard culture followed by drug susceptibility testing is still recommended[15]. While the Xpert MTB/RIF assay has the advantage of speed, it only looks at a specific small set of mutations associated with rifampicin resistance[39]. This has the potential to miss novel or rare variants that have not been included in the assay and does not detect resistance to other drugs. Using whole genome sequencing it is possible to characterise all variation in an isolate[21].

1.6 *Mycobacterium tuberculosis and strain diversity*

TB is caused by members of the *Mycobacterium tuberculosis* complex (MTBC, see **Figure 4** (top) for its position within the Mycobacterium phylogeny). The first whole genome sequence was published in 1998 by Cole *et al*[40]. The complex is characterised by low overall genetic diversity and a striking clonal population structure. *M. tuberculosis sensu stricto* consists of seven lineages; 1 Indo-Oceanic, 2 East-Asian including Beijing, 3 East-African-Indian, 4 Euro-American, 5 West African 1, 6 West African 2 and 7 Ethiopian[41]. These strains, together with *M. bovis* and other tuberculosis-causing animal strains make up the MTBC. Imperative in the study of infectious disease is the ability to compare the genetic relatedness of clinical strains of the pathogen of interest. This inference can be used to infer transmission dynamics and identify recent or ongoing outbreaks using sufficiently high resolution typing methods[42,43].

Several methods have been used to type (or genotype) *M. tuberculosis* (*Mtb*) including insertion sequence (IS6110) typing, spoligotyping and Mycobacterial interspersed repetitive units-variable number tandem repeat typing (MIRU-VNTR). The *Mtb* genome contains many insertion elements. The IS6110 insertion element has proven to be a good candidate to type as it is specific to members of the MTBC and varies in copy number[44]. Digestion using PvuII restriction enzyme cuts the DNA at specific sites in IS6110 sequence and when visualised using a southern blotting approach leads to a distinct fingerprint which differs by strain. Following this, PCR techniques were applied designed and greatly decreased the amount of DNA and time needed. The first PCR technique was named spoligotyping and made use of the direct repeat (DR) region in the *Mtb* genome[45]. The DR region contains a variable number of 36bp repeats interspersed by unique “spacer” sequences. The presence or absence of 43 of these spacer sequence differentiates between strain types. While having the advantage of needing a much smaller quantity of starting DNA than with IS6110 typing, the resolution is much lower and it should be using in conjunction with a high-resolution method. Another PCR based approach that has gained popularity is the MIRU-VNTR method[46]. This involved the analysis of variable number of tandem repeats (VNTR) of 24 repetitive loci in the *Mtb* genome including the mycobacterial interspersed repetitive units (MIRU)[47]. The loci are amplified and the number of repeats at each locus is estimated leading to a unique barcode which can be translated to a strain type. Using these methods many different strain types have been defined and deposited into databases such as SITVIT[48] and SpolDB4[49].

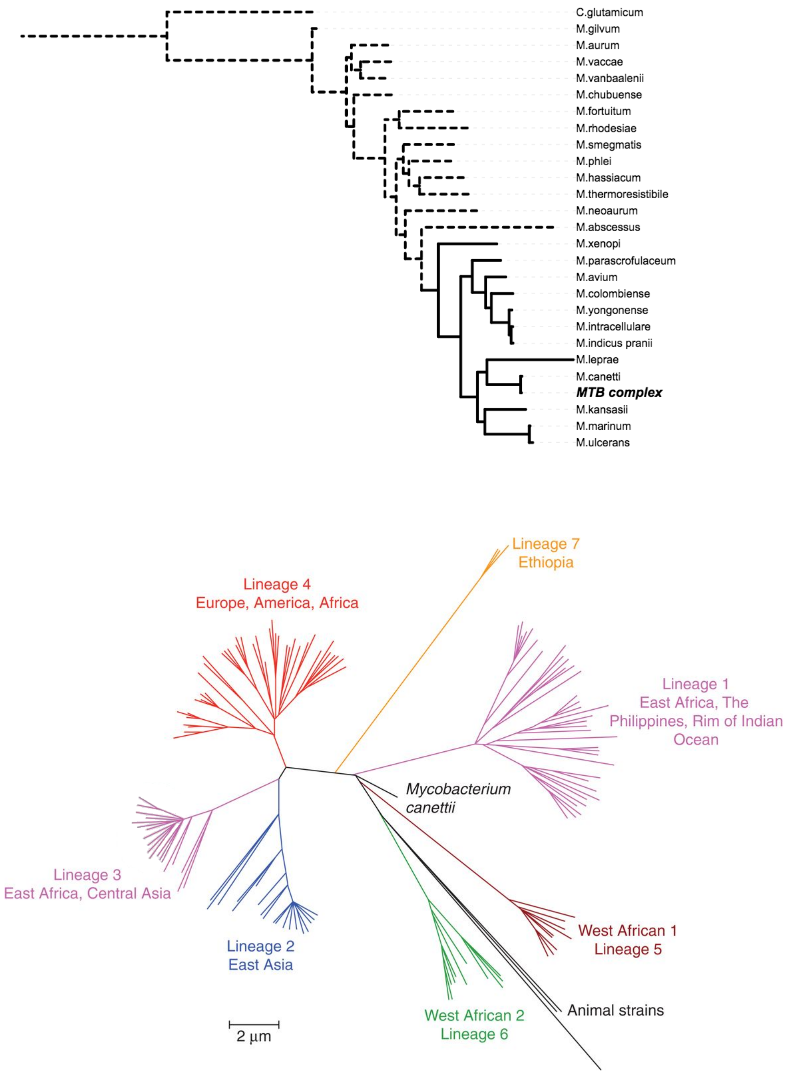
These genotyping methods use less than 1% of the *Mtb* genome (size: 4.4Mbp). *In silico* methods have also been developed to profile spoligotypes from whole genome sequence data[50]. Using these typing methods in conjunction with analysing long sequence polymorphisms (LSPs) or regions of difference (RDs) and whole genome sequencing approaches the phylogeny of the MTBC was delineated. Distinct clustering of strains into the seven lineages in a phylogenetic tree is expected (see **Figure 4** (bottom))**.** The strains have been designated “ancient” (lineage 1,5,6), “modern” (lineages 2,3 and 4) and intermediate (lineage 7) according to where in the tree they diverged from the ancestral outgroup (*M. canetti*). Studies on the phylogeographic spread of *Mtb* have shown a strong correlation between the strain type and geographic location[51]. This observation has led to the hypothesis that *Mtb* travelled with and co-evolved with their human hosts during the out of Africa expansion[52] . The lineages are postulated to have differential impacts on pathogenesis, disease outcome and vaccine efficacy[53–56]. For example, modern lineages, such as Beijing (lineage 2) and Euro-American Haarlem (lineage 4) strains exhibit more virulent phenotypes compared to ancient lineages, such as Indo-Oceanic[57]. Whilst some genetic differences between lineages have been identified[41], the molecular mechanisms responsible for differences in pathogenesis and virulence remain largely unknown[58].

1.7 *PE/PPE protein families*

Until recently, the *pe/ppe* genes have been difficult to sequence and have often been ignored as repetitive and potentially redundant gene families. However, these two groups of proteins, the pe and ppe families have recently been implicated in immune evasion and virulence[60]. Members of the *pe*/*ppe* gene families are characterised by the presence of proline-glutamate (pe) and proline-proline-glutamate (ppe) signature motifs near the N-terminus of their gene products[61]. The *pe* (99 loci) and *ppe* (69) gene families constitute ~10% of the coding potential of *M. tuberculosis* and are scattered throughout the genome[60]. The families can be subdivided based on similarities in their N-terminal regions[62]. Many of the *pe* and *ppe* gene products are predicted to be localised to the cell membrane or secreted including those in the *pe_pgrs* domain containing subgroup and the *ppe_mptr* domain containing subgroup[61,63]. It has been speculated that these proteins may play a role in virulence[61]. *Pe/ppe* genes are differentially expressed during infection[64] and some pe/ppe proteins have been shown to elicit immune responses by the host[61,65] and there is evidence that the pgrs domain can inhibit antigen processing[66]. Whilst *pe_pgrs* and *ppe_mptr* genes represent some of the most variable *M. tuberculosis* regions, some members of the *pe/ppe* family are conserved across strains and species, therefore implying different functional roles.

Figure 4 (top)

The mycobacterium phylogeny built using 27 whole genome reference sequences. The *M. tuberculosis* complex is located next to *M. canetti* in the slow growing mycobacterium clade (Phelan *et al*, 2015)[59]; (bottom) Phylogenetic tree depicting the main lineages of the MTBC. Adapted from Niemann *et al*[58]



Few of the pe/ppe protein structures (including pe25/ppe41) have been characterised[67], and *in lieu* of experimental and functional work, insights into their function and interaction partners must come from *in silico* analysis of large-scale 'omics data. However, due to the repetitive nature and high GC content genetic variation in the *pe/ppe* genes, it has been difficult to characterise them using traditional mapping approaches, leading to their systematic exclusion from analysis[67]. There have been conflicting studies reporting either high or little or no sequence divergence[68–70], but these studies have been limited by the number of genes and diversity of strains analysed.

1.8 *Next generation sequencing technologies*

Advances in genome sequencing technology have enabled the characterisation of the entire DNA sequence of an organism of interest. Next generation sequencing (NGS) refers to the high throughput sequencing technologies which superseded Sanger sequencing. Numerous NGS platforms have been developed including Illumina, 454 and Ion Torrent[71,72]. These platforms use a different set of reactions/processes however they all rely on the same principle: the DNA/genome is cut into smaller fragments and sequenced in parallel to produce a large number of overlapping sequences (reads). These sequences can be aligned to a reference genome or assembled *de novo* to build up a picture of the DNA which was sequenced. NGS has numerous advantages over sanger sequencing including cost, throughput and accuracy. The human genome project cost $3 billion and took 13 years to complete. With NGS it is now possible to sequence a human genome in three days for $1000[73]. Pathogen genomes are much smaller and multiple genomes can be sequenced ("multiplex sequencing") in parallel – thus driving the cost to ~£60 per isolate (based on an Illumina HighSeq, multiplexing 24 samples, 50-fold genomic coverage). The reduction in cost and increase in accuracy means that it is possible to sequence hundreds of genomes to use in projects investigating pathogen genomic diversity. To date, sequence data for ~31k isolates have been deposited into the European nucleotide archive (ENA) short reads archive. Whilst NGS technologies have provided significant improvements over previous technologies they do suffer from some drawbacks. The main limitation is that the short reads they produce do not characterise repetitive regions as well as unique regions of the genome. When the size of a repeat is longer than the read length it is extremely difficult to determine the copy number of a repeat using sequence information alone. Advancing on short-read sequencing technologies, so called third-generation NGS platforms have tried to circumvent these issues. These include Pacific Biosciences (PacBio)[74] and Oxford Nanopore minION[75]. Both these platforms have the ability to sequence much longer fragments, leading to the production of reads greater than 10Kb. The reads are long enough to span entire repetitive elements of *Mtb*. PacBio technology relies on sequencing of a template strand by a modified DNA polymerase. This process produces an optical signal which can be translated into a nucleotide base. Additionally, PacBio captures the speed at which a base is incorporated. DNA that has been methylated will take longer to pass through the polymerase. This differential speed allows for the epigenetic modifications to be detected. The minION uses a different technology, instead measuring the electrical current changes as a DNA molecule passes through a small pore in a membrane to identify the nucleotide bases. The minION is very portable and connects to a computer via USB connection to transfer the signal data which is then converted to base calls. Its portability and ease of use has let to its use in projects where mobility is key[76]. Although sequencing costs have dropped, the size of the genome means it is still not cost effective to run on a large scale for human studies such as GWASs. These studies use single nucleotide polymorphism (SNP) arrays to identify genetic differences, typically between TB cases and controls, whilst controlling for the confounding effects of population structure. The SNP arrays currently consist of millions of oligonucleotide probes (for example Illumina Omni 2.5), which hybridise selectively to DNA containing specific alleles. Additional genotype data up to ten million SNPs are imputed[77] from reference panels[78].

1.9 *Applications of whole genome sequencing*

The advent of NGS has enabled the characterisation of genomic variation at an ever-faster scale. This has allowed for numerous improvements in the classification of pathogen strains, detection of drug resistance and the large-scale study of transmission. NGS can detect differences between samples on a single base resolution. This fine-scale has made it an excellent choice of tool in the field of strain typing. Previous technologies such as IS6110, spoligotyping and MIRU-VNTR typing suffer from low resolution and convergent evolution of the same pattern. Whole genome sequencing on the other hand incorporates all possible genomic variation and thus provides much better resolution[58]. Multiple efforts have been made to create SNP barcodes that infer strain type from NGS data[79–81]. The largest of such studies was performed by Coll et al.[41] which analysed a collection of 1601 genomes resulting in a barcode of 62 SNPs. A numerical based lineage system was proposed allowing for nested sub-lineages. This system has allowed for the rapid classification of NGS data into strain types is useful in terms of epidemiological studies[82,83]. With the extra resolution NGS provides over other typing methods it has become possible to create transmission networks using genomic data along or in combination with epidemiological data. Mutations are acquired over time and the number of mutations between two isolates can be used as a proxy if a transmission event occurred. For *Mtb* a SNP a maximum of 10 SNP differences has been proposed as the cut-off for recent transmission[83,84]. NGS can also help to disentangle the origin of drug resistance in a patient; acquired or transmitted. Acquired drug resistance refers to the micro evolution within a patient that leads to the acquisition of drug resistance mutations. Transmitted resistance is the transmission of a strain which has already developed resistance to a drug. Studies have found a high level of transmitted resistance in high endemic regions such as South Africa[85,86]. Using NGS to create transmission networks and tracking the flow of drug resistance mutations through the network can shed insights on this topic. Another useful application of NGS is the *in silico* detection of drug resistance. Molecular diagnostics such as the GeneXpert MTB/RIF can detect specific mutations in the *rpoB* gene which lead to rifampicin resistance. While reporting high sensitivity, this could miss mutations which are not included in the assay. Using NGS it is possible to perform *in silico* resistance prediction from the resulting data, by detecting all variations and cross referencing these to a mutation database. Several computer programs providing fast and accurate prediction of drug resistance have been developed including TBProfiler, Mykrobe-predictor, PhyResSE[21,87,88]. One limitation with these tools is the underlying database. As sequencing becomes more common, new drug resistance variants will be detected or characterised, which have not been included in current databases. Recent consortial efforts are attempting to establish new databases based on whole genome sequence data and well characterised resistance phenotypes (ReSeqTB)[89]. Sequencing projects aiming to detect novel drug resistant variants are of great importance and will improve the diagnostic sensitivity of these tools. NGS can be envisioned running along standard diagnostics, helping to perform rapid *in-silico* drug resistance profiling to inform drug choices and identifying transmission events. Examples of its use in a clinical setting have already been demonstrated[90,91] and it will become more achievable with improvements in databases, software and the introduction of portable sequencing machines such as the minION.

1.10 *NGS analysis*

Raw sequence data and its associated quality from next generation sequencing machines typically is stored in text files called fastq files. The quality for each base is stored in the form of a phred quality score. The phred quality score represents the probability that the called base is incorrect. The raw sequence data can either be aligned to a reference genome (mapping) or assembled *de novo*. Mapping is used when fast accurate characterisation of non-repetitive regions is required. A multitude of programs are available to perform mapping including BWA and bowtie[92,93]. The mapping process consists of finding the optimum alignment position for each read in the dataset. The information is then stored in the SAM/BAM format[94]. SAMtools/BCFtools software can then be used to process the alignment files and extract SNPs and small indels. *De novo* assembly is used when a reference genome is not available or if there are hypervariable or repetitive regions in the genome. Programs such as Velvet[95] and SPAdes[96] can be used to perform assembly and output a fasta formatted sequence. This can then be aligned to a reference sequence or compared with other draft assemblies to extract variants. Variants and their associated qualities are stored in variant call files (VCF). A filtering step is required to take only high quality variants to downstream analyses. Variants with high quality score (Q>23), minimum depth of 10 and greater than 70% of reads supporting the variant are used in the final dataset. Variants for all isolates in a dataset can be collated together in a large matrix where the rows are variants and the columns are samples. This dataset can then be used in further population-level analyses. While NGS approaches works well with unique regions of the genome, a number of issues associated with the different steps during the data processing can lead to spurious variants. Firstly, the quality of the data has a large impact on the downstream results. Inclusion of low quality sequence can lead to the calling of false variants or aligning of a read to the wrong location. To reduce these errors, the raw data can be “trimmed” to remove bases with low quality phred scores. Secondly the non-unique regions of the genome, such as those coding for a domain which is present in many proteins, can lead to mapping of sequence to the wrong location which may lead to false variants. Either removing these regions or using *de novo* assembly can be used to counter these effects.

1.11 *Host-pathogen interactions*

TB can be viewed as a hierarchical model of two phenotypes interacting, that of the human and pathogen. These phenotypes are the result of a vast number of proteins, lipids and carbohydrates interacting together. In turn these proteins are coded for by the DNA and any variation in this sequence will have a knock-on effect on the phenotype. Though this representation is over simplified and disregards environmental and other stochastic processes, there is no doubt that variation in the genome influences the phenotype presented. For example, at an individual pathogen resolution there is evolution of drug resistance, and at a population scale, transmission dynamics of a strain and clonal outbreaks. The main theme of this work is to analyse variation in the host and pathogen to provide insights into phenotypes. Whilst this is mainly genomic variation with regards to drug resistance, it gets more complicated when analysing phenotypes such as clinical outcome when contact with the host is involved. The variability of the host could affect the variation of the bacterium and vice versa. The bacterium must be able to resist a wide range of environmental conditions, from airborne dehydration, macrophage uptake, endosome pH and immune attack. One possible route is to modulate the human host response to reduce the *Mtb* resistance to these environments. A recent study found differential expression of several genes in the *dosR* regulon between HIV positive and negative patients[97]. The function of the *dosR* regulon has not been fully characterised but it has been proposed to be involved with survival in granulomas[97]. The *pe*/*ppe* gene families have also been reported to elicit immune response from T and B cells and have been proposed to contribute towards host-evasion through antigenic variation[98,99].

1.12 The project structure

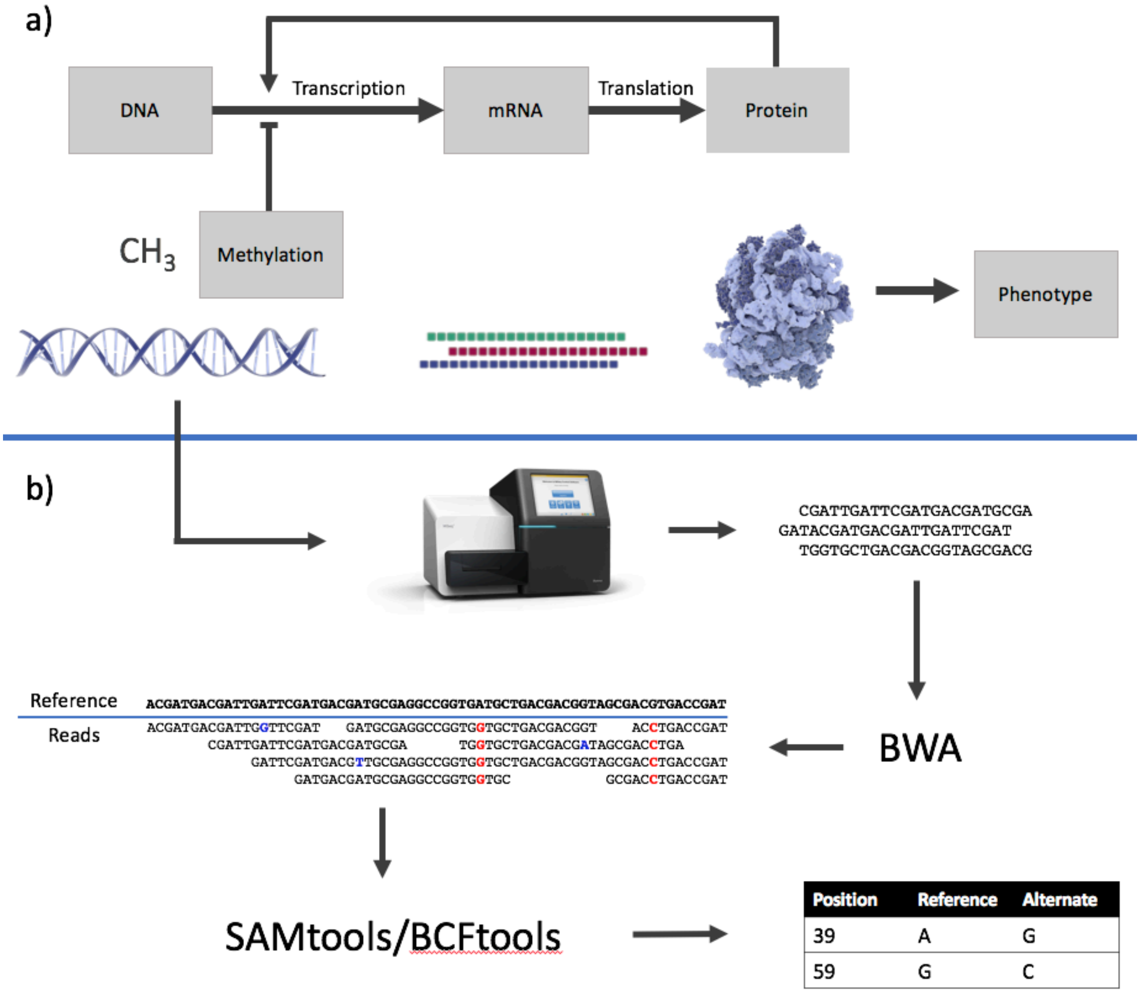
The overarching theme of this thesis is to use genetic information to improve our understanding of the impact of genetic variation on phenotypic traits such as drug resistance and host susceptibility to infection. **Figure 5** shows the

The thesis is divided it into 7 chapters, each consisting of one manuscript (4 published, 1 accepted, 1 under revision, and 1 in preparation), which address the following topics:

1. An evaluation of two sequencing platforms – the Illumina MiSeq and Ion torrent PGM – for sequencing in *Mtb*;
2. Validation of genome wide association studies to detect drug resistance mutations in *Mtb*;
3. Application of genome wide association study approach to a large global dataset of MDR and XDR strains;
4. Assembly of the first draft genome for *Mycobacterium aurum* – a surrogate model for anti-tuberculous drug screening;
5. Characterisation of the *pe/ppe* gene families using genome assembly in a set of ultra high-depth sequenced isolates;
6. Analysis of the *Mtb* methylome using PacBio sequencing;
7. Investigation into host-pathogen genome interactions using a GWAS approach.

Figure 5

This thesis analyses genomic data genomic, methylomic, protein structural and phenotypic data shown in (a). A simplified overview of the genomic data bioinformatics protocol is shown in (b).



```
CGATTGATTCGATGACGATGCGA
GATACGATGACGATTGATTCGAT
TGGTGCTGACGACGGTAGCGACG
```

```
Reference ACGATGACGATTGATTCGATGACGATGCGAGGCCGGTGATGCTGACGACGGTAGCGACGTGACCGAT
Reads     ACGATGACGATTGGTTCGAT   GATGCGAGGCCGGTGGTGCTGACGACGGT    ACCTGACCGAT
                 CGATTGATTCGATGACGATGCGA      TGGTGCTGACGACGATAGCGACCTGA
                      GATTCGATGACGTTGCGAGGCCGGTGGTGCTGACGACGGTAGCGACCTGACCGAT
                           GATGACGATGCGAGGCCGGTGGTGC           GCGACCTGACCGAT
```

Position	Reference	Alternate
39	A	G
59	G	C

Research papers included in this thesis in order of presentation include:

Research paper number (chapter)	Authors	Title	Status, journal and year
2	Phelan *et al* (including Clark TG)	The variability and reproducibility of whole genome sequencing technology for detecting resistance to anti-tuberculous drugs	Published. Genome Medicine (2016)
3	Phelan *et al* (including Clark TG)	Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance	Published. BMC Medicine (2016)
4	Phelan *et al* (including Hibberd ML and Clark TG)	The Mycobacterium tuberculosis resistome from a genome-wide analysis of multi- and extensively drug-resistant tuberculosis	In press. Nature Genetics
5	Phelan *et al* (including Bhakta S and Clark TG)	The draft genome of *Mycobacterium aurum*, a potential model organism for investigating drugs against *Mycobacterium tuberculosis* and *Mycobacterium leprae*.	Published. International journal of Mycobacteriology (2015)
6	Phelan *et al* (including Clark TG)	Recombination in pe/ppe genes contributes to genetic variation in *Mycobacterium tuberculosis* lineages	Published. BMC Genomics (2016)
7	Coll and Phelan *et al* (including Hibberd ML and Clark TG)	Methylation in Mycobacterium tuberculosis is lineage specific with associated mutations present globally	In press. Scientific reports
8	Coll and Phelan *et al* (including Hibberd ML and Clark TG)	Genome-wide host-pathogen analyses reveals genetic interaction points in tuberculosis disease	To be submitted. Scientific reports

The use of NGS to predict drug resistance relies on accurate characterisation of all sequence variants within an isolate. If the error rate is high or the sequence coverage is too low variants will be incorrectly called which will negatively impact on clinical decisions. Similarly, when inferring transmission, a high level of accuracy is required. Transmission is often inferred through measuring the number of SNP differences between isolates. A single error in variant calling could lead to a wrongly inferred transmission event. For large genome wide association studies, it is paramount to have a high-quality sequence dataset to boost true association signals. **Chapter 2** addresses the issue of variability and reproducibility of sequencing for use in a clinical and experimental setting. To this effect, we sequenced 10 M/XDR isolates and the reference strain (H37Rv) performing a number of technical and biological replicates in order to characterise the reproducibility of NGS for *Mtb*.

After establishing the high fidelity of sequencing and bioinformatic pipelines to process raw data to a set of high quality variant calls we proceeded to develop a robust genome wide association study pipeline. **Chapter 3** looks at the development of GWAS methods for drug resistance (DR) variant discovery. GWAS have been used in the human setting for over a decade but only recently have been applied in the prokaryotic field. I applied this methodology to a set of 127 drug resistant and sensitive clinical isolates from the TDR strain bank. In addition, we modelled the effect of mutations on protein structure and stability using crystallographic structures and homology models to identify key characteristics of DR mutations.

After validating the effectiveness of GWAS to detect drug resistant mutations we looked to apply the methodology to a larger dataset with potentially previously uncharacterised drug resistance mutations. **Chapter 4** describes the application of the GWAS methods detailed in **Chapter 3** to a large global collection of susceptible, MDR-TB and XDR-TB strains. We attempt to identify novel genetic variants to inform and improve *in silico* prediction of drug resistance.

Chapter 4 highlights the worrying amount of resistance to second line treatments. Whilst new drugs such as bedaquiline, linezolid and delamanid are being rolled out for use, resistance to these drugs has already been reported[100]. A steady flow of new or repurposed drugs are needed to combat the rise of XDR-TB. The high throughput profiling of therapeutic compounds in *Mtb* is hampered by its slow growth rate and high level of safety required. Several surrogate models have been proposed such as *M. smegmatis* and *M. fortuitum*. *M. aurum* is a fast-growing environmental mycobacterium which has proven to be useful as a model due to its similar cell wall composition and antibiotic susceptibility profiles. Whereas these phenotypic characteristics are comparable, their genomic similarities were not known due to the lack of reference genome for *M. aurum.* **Chapter 5** looks at assembling the draft genome and the genomic differences between *M. aurum* and *Mtb*.

In **Chapter 2**, I observed good coverage and high quality variant calling across loci involved in drug resistance. However, not all the regions of the genome are as easy to characterise. **Chapter 6** Looks at improving the characterisation of the *pe* and *pe* gene families, thought to play a role in host-pathogen interactions. These gene families are highly repetitive and are frequently omitted from population level analyses. By performing genome assembly using the previously developed pipeline (chapter 5) on a set of 518 high depth-of-coverage isolates I attempted to gain insights into the diversity within these families.

While **Chapters 1 to 6** focused on solely on genotype, there are additional modulating factors involved in the genotype to phenotype cascade, including transcription levels and methylation. Methylation has been studied in depth in human populations, giving rise to the field of epigenetics. Methylation can greatly influence a phenotype by altering gene expression whilst keeping the genetic code intact. Methylation also occurs in bacteria and whilst it mainly serves to protect its own DNA against restriction enzymes it has also been reported to play a role in transcription. In **Chapter 7,** I sought to characterise the extent that methylation occurs in *Mtb* by using PacBio sequencing. Additionally, long reads from PacBio enables a near-perfect characterisation of the repetitive regions in the genome including the *pe*/*ppe* genes, thus allowing confirmation of previous results.

When analysing aspects such as clinical outcome, host susceptibility and transmission, genomes from either the host or pathogen are analysed. In **Chapter 8**, I aim to consolidate both data sources into a single analysis. This approach allows us to look for co-occurrence of specific mutations in both genomes which may shed light on host-pathogen interactions. Previous attempts to detect susceptibility markers to tuberculosis have not been replicated across different human populations. We hypothesised that the differential endemic strains circulating could contribute toward this phenomenon. To test this hypothesis, I analysed samples from 720 tuberculosis positive patients for which we have human chip data and pathogen sequencing data, and revealed HLA – lineage interactions.

References

1. Organisation, W. H. Global tuberculosis report 2016. (2016).

2. Pagán, A. J. & Ramakrishnan, L. Immunity and Immunopathology in the Tuberculous Granuloma. *Cold Spring Harb. Perspect. Med.* **5,** a018499 (2015).

3. Stead, W. W., Senner, J. W., Reddick, W. T. & Lofgren, J. P. Racial differences in susceptibility to infection by Mycobacterium tuberculosis. *N. Engl. J. Med.* **322,** 422–7 (1990).

4. Cantwell, M. F., McKenna, M. T., McCray, E. & Onorato, I. M. Tuberculosis and Race/Ethnicity in the United States. *Am. J. Respir. Crit. Care Med.* **157,** 1016–1020 (1998).

5. Bellamy, R. Susceptibility to mycobacterial infections: the importance of host genetics. *Genes Immun.* **4,** 4–11 (2003).

6. Ehlers, S. & Schaible, U. E. The granuloma in tuberculosis: dynamics of a host-pathogen collusion. *Front. Immunol.* **3,** 411 (2012).

7. Medina, E. & North, R. J. Evidence inconsistent with a role for the Bcg gene (Nramp1) in resistance of mice to infection with virulent Mycobacterium tuberculosis. *J. Exp. Med.* **183,** 1045–51 (1996).

8. Hu, X., Peng, W., Chen, X., Zhao, Z., Zhang, J., Zhou, J., Cai, B., Chen, J., Zhou, Y., Lu, X. & Ying, B. No Significant Effect of ASAP1 Gene Variants on the Susceptibility to Tuberculosis in Chinese Population. *Medicine (Baltimore).* **95,** e3703 (2016).

9. Curtis, J., Luo, Y., Zenner, H. L., Cuchet-Lourenço, D., Wu, C., Lo, K., Maes, M., Alisaac, A., Stebbings, E., Liu, J. Z., Kopanitsa, L., Ignatyeva, O., Balabanova, Y., Nikolayevskyy,

V., Baessmann, I., Thye, T., Meyer, C. G., … Nejentsev, S. Susceptibility to tuberculosis is associated with variants in the ASAP1 gene encoding a regulator of dendritic cell migration. *Nat. Genet.* **47,** 523–527 (2015).

10. Miao, R., Ge, H., Xu, L., Sun, Z., Li, C., Wang, R., Ding, S., Yang, C. & Xu, F. Genetic variants at 18q11.2 and 8q24 identified by genome-wide association studies were not associated with pulmonary tuberculosis risk in Chinese population. *Infect. Genet. Evol.* **40,** 214–218 (2016).

11. Png, E., Alisjahbana, B., Sahiratmadja, E., Marzuki, S., Nelwan, R., Balabanova, Y., Nikolayevskyy, V., Drobniewski, F., Nejentsev, S., Adnan, I., van de Vosse, E., Hibberd, M. L., van Crevel, R., Ottenhoff, T. H. M. & Seielstad, M. A genome wide association study of pulmonary tuberculosis susceptibility in Indonesians. *BMC Med. Genet.* **13,** 5 (2012).

12. Thye, T., Vannberg, F. O., Wong, S. H., Owusu-Dabo, E., Osei, I., Gyapong, J., Sirugo, G., Sisay-Joof, F., Enimil, A., Chinbuah, M. A., Floyd, S., Warndorff, D. K., Sichali, L., Malema, S., Crampin, A. C., Ngwira, B., Teo, Y. Y., … Hill, A. V. S. Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. *Nat. Genet.* **42,** 739–741 (2010).

13. Khor, C. C. & Hibberd, M. L. Host–pathogen interactions revealed by human genome-wide surveys. *Trends Genet.* **28,** 233–243 (2012).

14. Lee, S.-W., Lin, C.-Y., Chuang, T.-Y., Huang, H.-H., Kao, Y.-H. & Wu, L. S.-H. SNP rs4331426 in 18q11.2 is associated with susceptibility to tuberculosis among female Han Taiwanese. *J. Microbiol. Immunol. Infect.* **49,** 436–438 (2016).

15. Zumla, A., Raviglione, M., Hafner, R. & von Reyn, C. F. Tuberculosis. *N. Engl. J. Med.*

368, 745–55 (2013).

16. Uddin, M. K. M., Chowdhury, M. R., Ahmed, S., Rahman, M. T., Khatun, R., van Leth, F. & Banu, S. Comparison of direct versus concentrated smear microscopy in detection of pulmonary tuberculosis. *BMC Res. Notes* **6,** 291 (2013).

17. Hermans, S. M., Babirye, J. A., Mbabazi, O., Kakooza, F., Colebunders, R., Castelnuovo, B., Sekaggya-Wiltshire, C., Parkes-Ratanshi, R. & Manabe, Y. C. Treatment decisions and mortality in HIV-positive presumptive smear-negative TB in the Xpert® MTB/RIF era: a cohort study. *BMC Infect. Dis.* **17,** 433 (2017).

18. Escalante, P. In the clinic. Tuberculosis. *Ann. Intern. Med.* **150,** ITC61-614; quiz ITV616 (2009).

19. Brown, A. C., Bryant, J. M., Einer-Jensen, K., Holdstock, J., Houniet, D. T., Chan, J. Z. M., Depledge, D. P., Nikolayevskyy, V., Broda, A., Stone, M. J., Christiansen, M. T., Williams, R., McAndrew, M. B., Tutill, H., Brown, J., Melzer, M., Rosmarin, C., … Breuer, J. Rapid Whole Genome Sequencing of M. tuberculosis directly from clinical samples. *J. Clin. Microbiol.* JCM.00486-15- (2015). doi:10.1128/JCM.00486-15

20. Witney, A. A., Gould, K. A., Arnold, A., Coleman, D., Delgado, R., Dhillon, J., Pond, M. J., Pope, C. F., Planche, T. D., Stoker, N. G., Cosgrove, C. A., Butcher, P. D., Harrison, T. S. & Hinds, J. Clinical application of whole-genome sequencing to inform treatment for multidrug-resistant tuberculosis cases. *J. Clin. Microbiol.* **53,** 1473–83 (2015).

21. Coll, F., McNerney, R., Preston, M., Guerra-Assunção, J. A., Warry, A., Hill-Cawthorn, G., Mallard, K., Nair, M., Miranda, A., Alves, A., Perdigão, J., Viveiros, M., Portugal, I., Hasan, Z., Hasan, R., Glynn, J. R., Martin, N., … Clark, T. G. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med.* **In**

Press, (2015).

22. Excellence, N. institute for health and care. NICE Guidelines [CG117]. (2011).

23. World Health Organization. *Global tuberculosis report 2014*. (2014).

24. Diacon, A. H., Pym, A., Grobusch, M., Patientia, R., Rustomjee, R., Page-Shipp, L., Pistorius, C., Krause, R., Bogoshi, M., Churchyard, G., Venter, A., Allen, J., Palomino, J. C., De Marez, T., van Heeswijk, R. P. G., Lounis, N., Meyvisch, P., … Neeley, D. F. M. The Diarylquinoline TMC207 for Multidrug-Resistant Tuberculosis. *N. Engl. J. Med.* **360,** 2397–2405 (2009).

25. Skripconoka, V., Danilovits, M., Pehme, L., Tomson, T., Skenders, G., Kummik, T., Cirule, A., Leimane, V., Kurve, A., Levina, K., Geiter, L. J., Manissero, D. & Wells, C. D. Delamanid improves outcomes and reduces mortality in multidrug-resistant tuberculosis. *Eur. Respir. J.* **41,** 1393–1400 (2013).

26. Gler, M. T., Skripconoka, V., Sanchez-Garavito, E., Xiao, H., Cabrera-Rivero, J. L., Vargas-Vasquez, D. E., Gao, M., Awad, M., Park, S.-K., Shim, T. S., Suh, G. Y., Danilovits, M., Ogata, H., Kurve, A., Chang, J., Suzuki, K., Tupasi, T., … Wells, C. D. Delamanid for Multidrug-Resistant Pulmonary Tuberculosis. *N. Engl. J. Med.* **366,** 2151–2160 (2012).

27. WHO | Treatment of drug-resistant TB: Resources. *WHO* (2017).

28. D’Ambrosio, L., Centis, R., Sotgiu, G., Pontali, E., Spanevello, A. & Migliori, G. B. New anti-tuberculosis drugs and regimens: 2015 update. *ERJ open Res.* **1,** (2015).

29. Nebenzahl-Guimaraes, H., Jacobson, K. R., Farhat, M. R. & Murray, M. B. Systematic review of allelic exchange experiments aimed at identifying mutations that confer

drug resistance in Mycobacterium tuberculosis. *J. Antimicrob. Chemother.* **69,** 331–42 (2014).

30. Almeida Da Silva, P. E. & Palomino, J. C. Molecular basis and mechanisms of drug resistance in Mycobacterium tuberculosis: classical and new drugs. *J. Antimicrob. Chemother.* **66,** 1417–1430 (2011).

31. Cohen, T., Sommers, B. & Murray, M. The effect of drug resistance on the fitness of Mycobacterium tuberculosis. *Lancet. Infect. Dis.* **3,** 13–21 (2003).

32. de Vos, M., Müller, B., Borrell, S., Black, P. A., van Helden, P. D., Warren, R. M., Gagneux, S. & Victor, T. C. Putative compensatory mutations in the rpoC gene of rifampin-resistant Mycobacterium tuberculosis are associated with ongoing transmission. *Antimicrob. Agents Chemother.* **57,** 827–32 (2013).

33. Sherman, D. R., Mdluli, K., Hickey, M. J., Arain, T. M., Morris, S. L., Barry, C. E. & Stover, C. K. Compensatory ahpC gene expression in isoniazid-resistant Mycobacterium tuberculosis. *Science* **272,** 1641–3 (1996).

34. White, R. J., Lancini, G. C. & Silvestri, L. G. Mechanism of action of rifampin on Mycobacterium smegmatis. *J. Bacteriol.* **108,** 737–41 (1971).

35. Ramaswamy, S. & Musser, J. M. Molecular genetic basis of antimicrobial agent resistance inMycobacterium tuberculosis: 1998 update. *Tuber. Lung Dis.* **79,** 3–29 (1998).

36. Lei, B., Wei, C. J. & Tu, S. C. Action mechanism of antitubercular isoniazid. Activation by Mycobacterium tuberculosis KatG, isolation, and characterization of inha inhibitor. *J. Biol. Chem.* **275,** 2520–6 (2000).

37. Heym, B., Alzari, P. M., Honoré, N. & Cole, S. T. Missense mutations in the catalase-peroxidase gene, katG, are associated with isoniazid resistance in Mycobacterium tuberculosis. *Mol. Microbiol.* **15,** 235–45 (1995).

38. Müller, B., Streicher, E. M., Hoek, K. G. P., Tait, M., Trollip, A., Bosman, M. E., Coetzee, G. J., Chabula-Nxiweni, E. M., Hoosain, E., Gey van Pittius, N. C., Victor, T. C., van Helden, P. D. & Warren, R. M. inhA promoter mutations: a gateway to extensively drug-resistant tuberculosis in South Africa? *Int. J. Tuberc. Lung Dis.* **15,** 344–51 (2011).

39. Chakravorty, S., Simmons, A. M., Rowneki, M., Parmar, H., Cao, Y., Ryan, J., Banada, P. P., Deshpande, S., Shenai, S., Gall, A., Glass, J., Krieswirth, B., Schumacher, S. G., Nabeta, P., Tukvadze, N., Rodrigues, C., Skrahina, A., … Alland, D. The New Xpert MTB/RIF Ultra: Improving Detection of Mycobacterium tuberculosis and Resistance to Rifampin in an Assay Suitable for Point-of-Care Testing. *MBio* **8,** e00812-17 (2017).

40. Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., … Barrell, B. G. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* **393,** 537–544 (1998).

41. Coll, F., McNerney, R., Guerra-Assunção, J. A., Glynn, J. R., Perdigão, J., Viveiros, M., Portugal, I., Pain, A., Martin, N. & Clark, T. G. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nat. Commun.* **5,** 4812 (2014).

42. Dong, H., Liu, Z., Lv, B., Zhang, Y., Liu, J., Zhao, X., Liu, J. & Wan, K. Spoligotypes of Mycobacterium tuberculosis from different Provinces of China. *J. Clin. Microbiol.* **48,** 4102–6 (2010).

43. Hasan, Z., Tanveer, M., Kanji, A., Hasan, Q., Ghebremichael, S. & Hasan, R. Spoligotyping of Mycobacterium tuberculosis isolates from Pakistan reveals predominance of Central Asian Strain 1 and Beijing isolates. *J. Clin. Microbiol.* **44,** 1763–8 (2006).

44. van Embden, J. D., Cave, M. D., Crawford, J. T., Dale, J. W., Eisenach, K. D., Gicquel, B., Hermans, P., Martin, C., McAdam, R. & Shinnick, T. M. Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. *J. Clin. Microbiol.* **31,** 406–409 (1993).

45. Groenen, P. M., Bunschoten, A. E., van Soolingen, D. & van Embden, J. D. Nature of DNA polymorphism in the direct repeat cluster of Mycobacterium tuberculosis; application for strain differentiation by a novel typing method. *Mol. Microbiol.* **10,** 1057–65 (1993).

46. Supply, P., Magdalena, J., Himpens, S. & Locht, C. Identification of novel intergenic repetitive units in a mycobacterial two-component system operon. *Mol. Microbiol.* **26,** 991–1003 (1997).

47. Supply, P., Allix, C., Lesjean, S., Cardoso-Oelemann, M., Rusch-Gerdes, S., Willery, E., Savine, E., de Haas, P., van Deutekom, H., Roring, S., Bifani, P., Kurepina, N., Kreiswirth, B., Sola, C., Rastogi, N., Vatin, V., Gutierrez, M. C., … van Soolingen, D. Proposal for Standardization of Optimized Mycobacterial Interspersed Repetitive Unit-Variable-Number Tandem Repeat Typing of Mycobacterium tuberculosis. *J. Clin. Microbiol.* **44,** 4498–4510 (2006).

48. Demay, C., Liens, B., Burguière, T., Hill, V., Couvin, D., Millet, J., Mokrousov, I., Sola, C., Zozio, T. & Rastogi, N. SITVITWEB – A publicly available international multimarker

database for studying Mycobacterium tuberculosis genetic diversity and molecular epidemiology. *Infect. Genet. Evol.* **12,** 755–766 (2012).

49. Brudey, K., Driscoll, J. R., Rigouts, L., Prodinger, W. M., Gori, A., Al-Hajoj, S. A., Allix, C., Aristimuño, L., Arora, J., Baumanis, V., Binder, L., Cafrune, P., Cataldi, A., Cheong, S., Diel, R., Ellermeier, C., Evans, J. T., … Sola, C. Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC Microbiol.* **6,** 23 (2006).

50. Coll, F., Mallard, K., Preston, M. D., Bentley, S., Parkhill, J., McNerney, R., Martin, N. & Clark, T. G. SpolPred: rapid and accurate prediction of Mycobacterium tuberculosis spoligotypes from short genomic sequences. *Bioinformatics* **28,** 2991–2993 (2012).

51. Hirsh, A. E., Tsolaki, A. G., DeRiemer, K., Feldman, M. W. & Small, P. M. Stable association between strains of Mycobacterium tuberculosis and their human host populations. *Proc. Natl. Acad. Sci.* **101,** 4871–4876 (2004).

52. Gagneux, S. Host-pathogen coevolution in human tuberculosis. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367,** 850–9 (2012).

53. Gagneux, S. & Small, P. M. Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. *Lancet. Infect. Dis.* **7,** 328–37 (2007).

54. de Jong, B. C., Hill, P. C., Aiken, A., Awine, T., Antonio, M., Adetifa, I. M., Jackson-Sillah, D. J., Fox, A., Deriemer, K., Gagneux, S., Borgdorff, M. W., McAdam, K. P. W. J., Corrah, T., Small, P. M. & Adegbola, R. A. Progression to active tuberculosis, but not transmission, varies by Mycobacterium tuberculosis lineage in The Gambia. *J. Infect. Dis.* **198,** 1037–43 (2008).

55. Caws, M., Thwaites, G., Dunstan, S., Hawn, T. R., Lan, N. T. N., Thuong, N. T. T., Stepniewska, K., Huyen, M. N. T., Bang, N. D., Loc, T. H., Gagneux, S., van Soolingen, D., Kremer, K., van der Sande, M., Small, P., Anh, P. T. H., Chinh, N. T., … Farrar, J. The influence of host and bacterial genotype on the development of disseminated disease with Mycobacterium tuberculosis. *PLoS Pathog.* **4,** e1000034 (2008).

56. Ordway, D. J., Shang, S., Henao-Tamayo, M., Obregon-Henao, A., Nold, L., Caraway, M., Shanley, C. A., Basaraba, R. J., Duncan, C. G. & Orme, I. M. Mycobacterium bovis BCG-mediated protection against W-Beijing strains of Mycobacterium tuberculosis is diminished concomitant with the emergence of regulatory T cells. *Clin. Vaccine Immunol.* **18,** 1527–35 (2011).

57. Reiling, N., Homolka, S., Walter, K., Brandenburg, J., Niwinski, L., Ernst, M., Herzmann, C., Lange, C., Diel, R., Ehlers, S. & Niemann, S. Clade-specific virulence patterns of Mycobacterium tuberculosis complex strains in human primary macrophages and aerogenically infected mice. *MBio* **4,** e00250-13 (2013).

58. Niemann, S. & Supply, P. Diversity and evolution of Mycobacterium tuberculosis: moving to whole-genome-based approaches. *Cold Spring Harb. Perspect. Med.* **4,** a021188 (2014).

59. Phelan, J., Maitra, A., McNerney, R., Nair, M., Gupta, A., Coll, F., Pain, A., Bhakta, S. & Clark, T. G. The draft genome of Mycobacterium aurum, a potential model organism for investigating drugs against Mycobacterium tuberculosis and Mycobacterium leprae. *Int. J. Mycobacteriology* **4,** 207–216 (2015).

60. Fishbein, S., van Wyk, N., Warren, R. M. & Sampson, S. L. Phylogeny to function: PE/PPE protein evolution and impact on Mycobacterium tuberculosis pathogenicity.

Mol. Microbiol. (2015). doi:10.1111/mmi.12981

61. Delogu, G., Cole, S. T. & Brosch, R. The PE and PPE Protein Families of Mycobacterium Tuberculosis. in *Handbook of Tuberculosis: Molecular Biology and Biochemistryitle* 131–150 (2008).

62. Gey van Pittius, N. C., Sampson, S. L., Lee, H., Kim, Y., van Helden, P. D. & Warren, R. M. Evolution and expansion of the Mycobacterium tuberculosis PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions. *BMC Evol. Biol.* **6,** 95 (2006).

63. Sampson, S. L., Lukey, P., Warren, R. M., van Helden, P. D., Richardson, M. & Everett, M. J. Expression, characterization and subcellular localization of the Mycobacterium tuberculosis PPE gene Rv1917c. *Tuberculosis (Edinb).* **81,** 305–17 (2001).

64. Delogu, G., Sanguinetti, M., Pusceddu, C., Bua, A., Brennan, M. J., Zanetti, S. & Fadda, G. PE_PGRS proteins are differentially expressed by Mycobacterium tuberculosis in host tissues. *Microbes Infect.* **8,** 2061–7 (2006).

65. Delogu, G. & Brennan, M. J. Comparative Immune Response to PE and PE_PGRS Antigens of Mycobacterium tuberculosis. *Infect. Immun.* **69,** 5606–11 (2001).

66. Singh, K. K., Zhang, X., Patibandla, A. S., Chien, P. & Laal, S. Antigens of Mycobacterium tuberculosis expressed during preclinical tuberculosis: serological immunodominance of proteins with repetitive amino acid sequences. *Infect. Immun.* **69,** 4185–91 (2001).

67. Galagan, J. E. Genomic insights into tuberculosis. *Nat. Rev. Genet.* **15,** 307–20 (2014).

68. Gutacker, M. M., Mathema, B., Soini, H., Shashkina, E., Kreiswirth, B. N., Graviss, E. A.

& Musser, J. M. Single-nucleotide polymorphism-based population genetic analysis of Mycobacterium tuberculosis strains from 4 geographic sites. *J. Infect. Dis.* **193,** 121–8 (2006).

69. Musser, J. M., Amin, A. & Ramaswamy, S. Negligible Genetic Diversity of Mycobacterium tuberculosis Host Immune System Protein Targets: Evidence of Limited Selective Pressure. *Genetics* **155,** 7–16 (2000).

70. Copin, R., Coscollá, M., Seiffert, S. N., Bothamley, G., Sutherland, J., Mbayo, G., Gagneux, S. & Ernst, J. D. Sequence diversity in the pe_pgrs genes of Mycobacterium tuberculosis is independent of human T cell recognition. *MBio* **5,** e00960-13 (2014).

71. Quail, M., Smith, M. E., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P. & Gu, Y. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics* **13,** 341 (2012).

72. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17,** 333–351 (2016).

73. Illumina. Available at: https://www.illumina.com.

74. Nakano, K., Shiroma, A., Shimoji, M., Tamotsu, H., Ashimine, N., Ohki, S., Shinzato, M., Minami, M., Nakanishi, T., Teruya, K., Satou, K. & Hirano, T. Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. *Hum. Cell* **30,** 149–161 (2017).

75. Jain, M., Tyson, J. R., Loose, M., Ip, C. L. C., Eccles, D. A., O'Grady, J., Malla, S., Leggett, R. M., Wallerman, O., Jansen, H. J., Zalunin, V., Birney, E., Brown, B. L., Snutch, T. P., Olsen, H. E. & MinION Analysis and Reference Consortium. MinION Analysis and

Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry. *F1000Research* **6,** 760 (2017).

76. Faria, N. R., Sabino, E. C., Nunes, M. R. T., Alcantara, L. C. J., Loman, N. J. & Pybus, O. G. Mobile real-time surveillance of Zika virus in Brazil. *Genome Med.* **8,** 97 (2016).

77. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10,** 387–406 (2009).

78. Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E. & McVean, G. A. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–73 (2010).

79. Filliol, I., Motiwala, A. S., Cavatore, M., Qi, W., Hazbón, M. H., Bobadilla del Valle, M., Fyfe, J., García-García, L., Rastogi, N., Sola, C., Zozio, T., Guerrero, M. I., León, C. I., Crabtree, J., Angiuoli, S., Eisenach, K. D., Durmaz, R., … Alland, D. 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. Bacteriol.* **188,** 759–72 (2006).

80. Comas, I., Homolka, S., Niemann, S., Gagneux, S. & Beckstrom-Sternberg, S. Genotyping of Genetically Monomorphic Bacteria: DNA Sequencing in Mycobacterium tuberculosis Highlights the Limitations of Current Methodologies. *PLoS One* **4,** e7815 (2009).

81. Homolka, S., Projahn, M., Feuerriegel, S., Ubben, T., Diel, R., Nübel, U. & Niemann, S. High Resolution Discrimination of Clinical Mycobacterium tuberculosis Complex Strains Based on Single Nucleotide Polymorphisms. *PLoS One* **7,** e39855 (2012).

82. Guerra-Assunção, J., Crampin, A., Houben, R., Mzembe, T., Mallard, K., Coll, F., Khan, P., Banda, L., Chiwaya, A., Pereira, R., McNerney, R., Fine, P., Parkhill, J., Clark, T. & Glynn, J. Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *Elife* **4,** (2015).

83. Tyler, A. D., Randell, E., Baikie, M., Antonation, K., Janella, D., Christianson, S., Tyrrell, G. J., Graham, M., Van Domselaar, G. & Sharma, M. K. Application of whole genome sequence analysis to the study of Mycobacterium tuberculosis in Nunavut, Canada. *PLoS One* **12,** e0185656 (2017).

84. Glynn, J. R., Guerra-Assunção, J. A., Houben, R. M. G. J., Sichali, L., Mzembe, T., Mwaungulu, L. K., Mwaungulu, J. N., McNerney, R., Khan, P., Parkhill, J., Crampin, A. C. & Clark, T. G. Whole Genome Sequencing Shows a Low Proportion of Tuberculosis Disease Is Attributable to Known Close Contacts in Rural Malawi. *PLoS One* **10,** e0132840 (2015).

85. Shah, N. S., Auld, S. C., Brust, J. C. M., Mathema, B., Ismail, N., Moodley, P., Mlisana, K., Allana, S., Campbell, A., Mthiyane, T., Morris, N., Mpangase, P., van der Meulen, H., Omar, S. V., Brown, T. S., Narechania, A., Shaskina, E., … Gandhi, N. R. Transmission of Extensively Drug-Resistant Tuberculosis in South Africa. *N. Engl. J. Med.* **376,** 243–253 (2017).

86. Dheda, K., Limberis, J. D., Pietersen, E., Phelan, J., Esmail, A., Lesosky, M., Fennelly, K. P., te Riele, J., Mastrapa, B., Streicher, E. M., Dolby, T., Abdallah, A. M., Ben-Rached, F., Simpson, J., Smith, L., Gumbo, T., van Helden, P., … Warren, R. M. Outcomes, infectiousness, and transmission dynamics of patients with extensively drug-resistant tuberculosis and home-discharged patients with programmatically incurable

tuberculosis: a prospective cohort study. *Lancet Respir. Med.* **5,** 269–281 (2017).

87. Bradley, P., Gordon, N. C., Walker, T. M., Dunn, L., Heys, S., Huang, B., Earle, S., Pankhurst, L. J., Anson, L., de Cesare, M., Piazza, P., Votintseva, A. A., Golubchik, T., Wilson, D. J., Wyllie, D. H., Diel, R., Niemann, S., … Iqbal, Z. Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nat. Commun.* **6,** 10063 (2015).

88. Feuerriegel, S., Schleusener, V., Beckert, P., Kohl, T. A., Miotto, P., Cirillo, D. M., Cabibbe, A. M., Niemann, S. & Fellenberg, K. PhyResSE: a Web Tool Delineating Mycobacterium tuberculosis Antibiotic Resistance and Lineage from Whole-Genome Sequencing Data. *J. Clin. Microbiol.* **53,** 1908–1914 (2015).

89. Starks, A. M., Avilés, E., Cirillo, D. M., Denkinger, C. M., Dolinger, D. L., Emerson, C., Gallarda, J., Hanna, D., Kim, P. S., Liwski, R., Miotto, P., Schito, M. & Zignol, M. Collaborative Effort for a Centralized Worldwide Tuberculosis Relational Sequencing Data Platform: Figure 1. *Clin. Infect. Dis.* **61,** S141–S146 (2015).

90. Arnold, A., Witney, A. A., Vergnano, S., Roche, A., Cosgrove, C. A., Houston, A., Gould, K. A., Hinds, J., Riley, P., Macallan, D., Butcher, P. D. & Harrison, T. S. XDR-TB transmission in London: Case management and contact tracing investigation assisted by early whole genome sequencing. *J. Infect.* **73,** 210–218 (2016).

91. Witney, A. A., Cosgrove, C. A., Arnold, A., Hinds, J., Stoker, N. G. & Butcher, P. D. Clinical use of whole genome sequencing for Mycobacterium tuberculosis. *BMC Med.* **14,** 46 (2016).

92. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).

93. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–9 (2012).

94. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–9 (2009).

95. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–9 (2008).

96. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V, Sirotkin, A. V, Vyahhi, N., Tesler, G., Alekseyev, M. A. & Pevzner, P. A. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19,** 455–77 (2012).

97. Walter, N. D., de Jong, B. C., Garcia, B. J., Dolganov, G. M., Worodria, W., Byanyima, P., Musisi, E., Huang, L., Chan, E. D., Van, T. T., Antonio, M., Ayorinde, A., Kato-Maeda, M., Nahid, P., Leung, A. M., Yen, A., Fingerlin, T. E., … Schoolnik, G. K. Adaptation of *Mycobacterium tuberculosis* to Impaired Host Immunity in HIV-Infected Patients. *J. Infect. Dis.* **214,** 1205–1211 (2016).

98. Brennan, M. J. The Enigmatic PE/PPE Multigene Family of Mycobacteria and Tuberculosis Vaccination. *Infect. Immun.* **85,** e00969-16 (2017).

99. Sampson, S. L. Mycobacterial PE/PPE Proteins at the Host-Pathogen Interface. *Clin. Dev. Immunol.* **2011,** 1–11 (2011).

100. Xu, J., Wang, B., Hu, M., Huo, F., Guo, S., Jing, W., Nuermberger, E. & Lu, Y. Primary Clofazimine and Bedaquiline Resistance among Isolates from Patients with Multidrug-

Resistant Tuberculosis. *Antimicrob. Agents Chemother.* **61,** e00239-17 (2017).

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Jody Phelan
Principal Supervisor	Taane Clark
Thesis Title	A Bioinformatic analysis of M. tuberculosis and host genomic data

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?	Genome Medicine		
When was the work published?	December 2016		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I received the raw fastq data from our collaborators. I designed and ran the analysis pipeline, consisting of read trimming, mapping and calling variants. I then designed custom python scripts to extract information on coverage and to assess concordance of the variants called between replicates. All figures were plotted using custom scripts which I wrote. I used the TBProfiler and Mykrobe TB predictor to predict drug resistance variants *in-silico* and tabulated all the results in excel. I wrote the first draft of the manuscript and circulated to co-authors. After several iterations of including comments from co-authors I uploaded to the Genome medicine submission portal and dealt with any subsequent revisions.

Student Signature: ____________________________________ **Date:** __________________

Supervisor Signature: ____________________________________ **Date:** __________________

Chapter 2

The variability and reproducibility of whole genome sequencing technology for detecting resistance to anti-tuberculous drugs

Phelan *et al. Genome Medicine* (2016) 8:132
DOI 10.1186/s13073-016-0385-x

Genome Medicine

RESEARCH Open Access

The variability and reproducibility of whole genome sequencing technology for detecting resistance to anti-tuberculous drugs

Jody Phelan[1†], Denise M. O'Sullivan[2†], Diana Machado[3†], Jorge Ramos[3], Alexandra S. Whale[2], Justin O'Grady[4], Keertan Dheda[5], Susana Campino[1], Ruth McNerney[5†], Miguel Viveiros[3†], Jim F. Huggett[2,6†] and Taane G. Clark[1,7*†]

Abstract

Background: The emergence of resistance to anti-tuberculosis drugs is a serious and growing threat to public health. Next-generation sequencing is rapidly gaining traction as a diagnostic tool for investigating drug resistance in *Mycobacterium tuberculosis* to aid treatment decisions. However, there are few little data regarding the precision of such sequencing for assigning resistance profiles.

Methods: We investigated two sequencing platforms (Illumina MiSeq, Ion Torrent PGM™) and two rapid analytic pipelines (*TBProfiler*, *Mykrobe predictor*) using a well characterised reference strain (H37Rv) and clinical isolates from patients with tuberculosis resistant to up to 13 drugs. Results were compared to phenotypic drug susceptibility testing. To assess analytical robustness individual DNA samples were subjected to repeated sequencing.

Results: The MiSeq and Ion PGM systems accurately predicted drug-resistance profiles and there was high reproducibility between biological and technical sample replicates. Estimated variant error rates were low (MiSeq 1 per 77 kbp, Ion PGM 1 per 41 kbp) and genomic coverage high (MiSeq 51-fold, Ion PGM 53-fold). MiSeq provided superior coverage in GC-rich regions, which translated into incremental detection of putative genotypic drug-specific resistance, including for resistance to para-aminosalicylic acid and pyrazinamide. The *TBProfiler* bioinformatics pipeline was concordant with reported phenotypic susceptibility for all drugs tested except pyrazinamide and para-aminosalicylic acid, with an overall concordance of 95.3%. When using the *Mykrobe predictor* concordance with phenotypic testing was 73.6%.

Conclusions: We have demonstrated high comparative reproducibility of two sequencing platforms, and high predictive ability of the *TBProfiler* mutation library and analytical pipeline, when profiling resistance to first- and second-line anti-tuberculosis drugs. However, platform-specific variability in coverage of some genome regions may have implications for predicting resistance to specific drugs. These findings may have implications for future clinical practice and thus deserve further scrutiny, set within larger studies and using updated mutation libraries.

Keywords: Drug resistance, Tuberculosis, Diagnostics, Drug-susceptibility testing, XDR-TB, Next-generation sequencing

* Correspondence: taane.clark@lshtm.ac.uk
†Equal contributors
[1]Department of Pathogen Molecular Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, WC1E 7HT London, UK
[7]Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, WC1E 7HT London, UK
Full list of author information is available at the end of the article

Background

Mycobacterium tuberculosis, the bacterium that causes tuberculosis disease (TB), has overtaken HIV as the world's major cause of death from an infectious agent [1]. In recent years, control of the disease has been made more difficult by the emergence of multidrug-resistant tuberculosis (MDR-TB), which is resistant to at least rifampicin and isoniazid, and extensively drug-resistant (XDR-TB), which refers to additional resistance to the fluoroquinolones and second-line injectable drugs (amikacin, kanamycin and capreomycin) used to treat MDR-TB [2]. Programmatically incurable TB with resistance to up to 14 drugs has been reported in several parts of the world, including countries with a high TB burden such as India and South Africa [3, 4]. Phenotypic methods of determining susceptibility to anti-TB drugs take weeks or months, they are additively costly, and require culture and manipulation of large numbers of highly infectious bacilli. Drug resistance in *M. tuberculosis* is almost exclusively due to mutations in the circular genome and so molecular determination of resistance offers a rapid, potentially cost effective, and safer alternative. Commercially available molecular-based tests and line probe assays cover a limited number of drugs but, with the exception of rifampicin, they have relatively low sensitivity for detecting all possible molecular targets for resistance [5]. Due to the multiplicity of drugs used in the treatment of TB, determining the full resistance profile for a patient suspected of having drug-resistant disease requires the examination of many loci.

Next-generation whole genome sequencing offers an attractive option as it simultaneously examines all loci and provides information regarding both small and large changes in the genome [5]. This option has been widely reported as a means of identifying putative resistance-causing mutations and more recently has been used in the management of patients with drug-resistant TB to guide selection of appropriate drug regimens [6–11]. This approach is significant because the current treatment outcomes for MDR-TB are poor, largely due to current molecular tests being unable to guide effective individualised therapy. It also has public health implications because of prolonged patient infectiousness due to suboptimal treatment.

The *M. tuberculosis* genome is challenging to sequence due to its high GC content and repetitive nature. Surprisingly, despite the serious consequences of misdiagnosis, there is a paucity of data regarding the reliability of next-generation sequencing platforms or the analytical methodology used for assigning resistance [5]. To address this issue we investigated the utility of two commercial sequencing platforms for predicting resistance to 13 anti-TB drugs. We also examined analytical algorithms and two rapid bioinformatics tools (*TBProfiler*, *Mykrobe predictor*) for predicting resistance from raw sequence data. Testing was performed with a fully susceptible reference strain (H37Rv) and ten clinical isolates from patients with drug-resistant TB.

Methods

Samples

M. tuberculosis clinical isolates were sourced from ten patients with known drug-resistant TB admitted to four different hospitals in Lisbon between 2007 and 2013. These samples were not part of a transmission chain and there is no epidemiological link between the patients. All clinical samples and the reference strain H37Rv (ATCC 25618D-9, Lot # 60986340) were prepared by inoculating a single colony into Middlebrook 7H9 broth supplemented with 10% OADC (Becton Dickinson) (see Table 1 for list). Susceptibility testing for the first-line anti-TB drugs rifampicin (RIF), isoniazid (INH), ethambutol (ETB), pyrazinamide (PZA) and streptomycin (STR) and the second-line drugs rifabutin (RFB), amikacin (AMK), capreomycin (CAP), ofloxacin (OFX), moxifloxacin (MOX), ethionamide (ETH), para-aminosalicylic acid (PAS) and linezolid (LZ) was performed on all strains with the MGIT960 system (Becton Dickinson), according to the manufacturer's instructions. Quantitative drug susceptibility testing (qDST) for both first- and second-line drugs was conducted using a combination of the MGIT960 system and the Epicenter V5.80A software equipped with the TB eXIST module (Becton Dickinson) [12, 13].

DNA was extracted and purified from the liquid cultures using a cetyltrimethylammonium bromide (CTAB) method [14]. The quality was assessed by fluorometric quantification, Qubit™ 3.0 Fluorometer with a dsDNA Broad Range Assay Kit (Thermo Fisher Scientific) and agarose gel electrophoresis. Triplicate DNA samples from each clinical isolate were prepared (biological replicates) and individual DNA extracts were subjected to repeated sequencing (technical replicates).

Library preparation and sequencing

For MiSeq sequencing, ~200 ng of genomic DNA was sheared to an average size of 500 bp by ultrasonication (Covaris S220). Sheared DNA was purified/concentrated on MinElute Spin Columns (Qiagen). DNA concentrations were measured on a Nanodrop UV spectrophotometer and the sheared samples diluted to 5–12.5 ng/μl. Library constructions were performed using the Ovation Rapid DR Multiplex System (NuGen) according to the manufacturer's instructions. Purified libraries were amplified in emulsion PCR, size selected (500–700 bp) by preparative electrophoresis on composite gels (1.2% LMP-Agarose/0.8% Synergel) and then purified on MinElute Columns. Libraries were sequenced

Table 1 Study samples (DNA extracted from culture isolates) and their susceptibility to anti-tuberculosis drugs

Sample	Year[a]	Lineage	Spoligo. family	Drug susceptibility test phenotype														
				INH	RIF	STR	ETB	PZA	RFB	ETH	AMK	CAP	OFX	MOX	PAS	LZ	KAN[b]	Resistance phenotype
POR1	2007	4.3.4.2	LAM4	R	R	**R**	R	R	R	R	R	R	R	R	$\underline{\text{R}}$	S	R	XDR-TB
POR2	2007	4.1.1.1	X2	**R**	R	S	S	S	R	R	S	S	S	S	S	S	-	MDR-TB
POR3	2007	4.3.4.2	LAM1	R	R	R	**R**	$\underline{\text{R}}$	R	**R**	**R**	**R**	R	R	S	S	**R**	XDR-TB
POR4	2007	4.3.4.2	LAM1	R	R	R	R	R	R	R	**R**	S	R	R	S	S	**R**	XDR-TB
POR5	2007	4.3.4.2	LAM4	R	R	**R**	R	R	R	R	S	S	S	S	S	S	-	MDR-TB
POR6	2008	4.3.4.2	LAM4	R	R	**R**	R	R	R	R	R	R	R	R	S	S	R	XDR-TB
POR7	2009	4.3.4.2	LAM4	R	R	R	R	R	R	R	R	R	R	R	S	S	**R**	XDR-TB
POR8	2012	4.3.4.2	LAM4	R	R	**R**	R	R	R	R	R	R	R	R	S	S	R	XDR-TB
POR9	2011	4.3.4.2	LAM4	R	R	R	**R**	R	R	**R**	**R**	**R**	R	R	$\underline{\text{R}}$	S	**R**	XDR-TB
POR10	2013	4.2.1	Ural H3/4	R	R	R	R	$\underline{\text{R}}$	R	R	S	S	S	S	S	S	**R**	MDR-TB
H37Rv	-	4.9	H37RV	S	S	S	S	S	S	S	S	S	S	S	S	S	-	Pan-susceptible

MDR-TB multidrug-resistant TB, *XDR-TB* extensively drug-resistant TB, *INH* isoniazid, *RIF* rifampicin, *STR* streptomycin, *ETB* ethambutol, *PZA* pyrazinamide, *RFB* rifabutin, *ETH* ethionamide, *AMK* amikacin, *CAP* capreomycin, *OFX* ofloxacin, *MOX* moxifloxacin, *PAS* para-aminosalicylic acid, *LZ* linezolid, *KAN* kanamycin, *S* "susceptible", *R* "resistant"

Bold indicates discrepant calls by *Mykrobe Predictor*, underlining indicates discrepant calls by *TBProfiler*

[a] Year of collection

[b] Drug susceptibility test not performed, with status inferred by the *TBProfiler* library

with an Illumina MiSeq V3 and 300-bp paired-end reads with samples randomised across two runs (each ~24 h in duration).

Ion Torrent library preparation and sequencing was performed at Thermo Fisher Scientific, UK. Sequencing was carried out with the Ion Torrent PGM™ system (Ion PGM). Libraries were constructed with the Ion Xpress™ Plus Fragment Library Kit as per the manufacturer's instructions (MAN0009847 Revision C.0), using 100 ng of genomic DNA which was sheared with the provided Ion Shear™ Plus Reagents to an average size of 350 bp, size selection using an E-Gel® SizeSelect™ 2% Agarose Gel, and purification with Agencourt® AMPure® XP Reagent. Finally, the libraries were quantified on the StepOnePlus™ System using the Ion Library Quantitation Kit, then diluted to 100 pM and pooled in equal volume. Purified libraries were sequenced with an Ion 318™ v2 chip (400-bp kit) and the Ion PGM™ HiQ™ Chef Kit as stated in the manual (MAN0010919, revision A.0). The runtime was ~3 h per sample. The software used on both Ion PGM™ and the Ion Chef™ System was Torrent Suite™ Software version 4.6.

Bioinformatic pipeline

For the bioinformatic analysis we used a previously reported pipeline [10, 15, 16]. Unless stated otherwise, software was run at default settings. Reads were trimmed by *Trimmomatic* using a PHRED quality of 20 as the cutoff. Trimmed reads were then mapped against H37Rv (GCA_000195955.2) with *BWA-mem* v0.7.12 [17]. SNP and insertion and deletion (indel) variants were called with *Samtools* 0.1.19 [18] and *GATK* v3.6 [19]. We compared the variants called by both algorithms, but also report results of the conservative and typical approach of retaining the consensus polymorphisms across both methods. The genotypes of SNPs were called when an alternative allele was found in 20% of the mapped reads at a particular position. A default minimum depth of ten reads was required to call SNP genotypes, otherwise genotypes were denoted as missing data. This cutoff has been applied widely [15, 16, 20]. The robustness of the genotype calls was assessed across a range of depths of coverage of the reference and alternative alleles (depth 5–20, major allelic frequency >0.5 and >0.7). The reference genome was partitioned into overlapping 300-bp sequences allowing the uniqueness of genomic regions to be determined using *gem-mappability* [21]. Only 1.5% of the genome was estimated to be non-unique, and variants within these regions were discarded, leaving a set of high quality SNPs and indels. All 36 candidate drug-resistance genes [5] were found to be unique, thus removing the risk of false calling of SNPs due to inappropriate mapping to an analogous region. A summary of the pipeline is presented in Additional file 1: Figure S1.

In silico profiling of *M. tuberculosis* resistance phenotypes

We compared two informatics tools for assigning resistance from sequence data. Drug-resistance status across 14 drugs was called in silico from raw sequence data using the web-based *TBProfiler* tool (http://tbdr.lshtm.ac.uk/) [5]). This tool also generates lists of mutations in candidate loci, and these formed the basis of identifying any additional putative novel polymorphisms. All mutations were checked by analysis of alignments and de novo assembly, as well as confirmed by alternative sequencing methods (see the next section, "Confirmation of mutations detected by whole genome sequencing"). Resistance profiles were also generated with the *Mykrobe predictor* tool (version July 2016) [22].

Confirmation of mutations detected by whole genome sequencing

Genomic DNA was extracted as described above and used for PCR amplification prior to examination by line probe assay and/or DNA sequencing. The Genotype MTBDR*plus* (Hain Lifescience) investigates the *rpoB* and *katG* genes and *inhA* regulatory region and Genotype MTBDR*sl* (version 1, Hain Lifescience) investigates *rrs*, *gyrA* and *embB*. Both kits were used according to the manufacturer's instructions. As the line probe assays encompass a limited number of loci, we also performed Sanger sequencing for *inhA*, *katG*, *tlyA*, *eis*, *gidB*, *pncA*, *gyrA*, *ethA*, *embB*, *embC-embA*, *rpsL*, *folC* and *thyX* genes (see Additional file 2: Table S1 for the primers used). PCR products were purified and both strands sequenced at StabVida (Portugal). All sequences were edited and analysed with ChromasPro 2.0.0 (Technelysium, Australia), compared to the sequences of *M. tuberculosis* H37Rv reference strain (GenBank AL123456.2) and aligned with Clustal Omega [23].

Results

Coverage

Triplicate "extraction" DNA samples from ten clinical isolates and a single H37Rv sample were sequenced on the MiSeq platform. Four DNA samples (from POR5, 6 and 7 and H37Rv) were each sequenced six times ("technical" replicates). Duplicate DNA samples from three clinical isolates (POR1, 2 and 6) were also sequenced on the Ion PGM. Summaries of the sequence data obtained for each platform are presented in Additional file 3: Table S2. With MiSeq sequencing the number of paired reads varied across samples (median 1.2 million, range 0.4 to 3.2 million), and on average 99% of reads mapped to the H37Rv reference, giving a median depth of coverage of 51-fold (across sample range 18- to 79-fold). The majority of the genome (>96%) was covered to at least tenfold depth.

In contrast, for the Ion PGM platform the median number of reads was 990,854 (range 928,006–1,124,215) translating into a median of 53-fold (range 48- to 59-fold) genomic coverage. A large proportion of the genome (~25%) had low coverage and was attributed to regions with high GC content (Fig. 1). Whilst high coverage (100- to 200-fold) was attained for regions with GC content up to 69%, above this level coverage drops below tenfold, which was the cutoff used for calling variants. For MiSeq sequence data, this drop only occurs when the GC content reaches 75% or above. Many regions in the *M. tuberculosis* genome, especially the *pe/ppe* genes [24], are high in GC content (median 69%, range 47–87%) and therefore potentially difficult to characterise. The coverage across the 36 drug-resistance candidate genes was high for MiSeq (mean ~90-fold) and exceeded the tenfold cutoff, except in the *thyA* gene in the three POR1 replicates (Fig. 2). These XDR-TB replicates contained double *dfrA-thyA* deletions, thought to be responsible for para-aminosalicylic acid (PAS) resistance [25]. A direct comparison of the POR1, 2 and 6 sample coverage across platforms highlighted greater variability in candidate genes in Ion PGM due to differential GC content. Whilst there was platform-wide detection of the deletion-driven lower coverage in *thyA* in POR1 (Fig. 3; Additional file 4: Figure S2), the variable coverage in the neighbouring regions for Ion PGM could lead to less certainty in detection.

SNP variants and error rates

We estimated the variant error rates (measured as the number of sites which were discordant among replicates) to be low for both platforms (MiSeq 1 per 77 kbp, Ion PGM 1 per 41 kbp). Across comparable samples, the number of high quality SNPs detected using MiSeq data was higher than from Ion PGM, mostly due to low coverage in the alignments generated from the Ion PGM (Additional file 3: Table S2). We sought to investigate the effects of variant calling algorithms on the numbers of SNPs detected in unique genomic regions. From the MiSeq H37Rv data, similar numbers of SNPs were detected across replicates (*Samtools* 64–69 SNPs and *GATK* 69–79 SNPs, overlap 69 SNPs), supporting the existence of those variants and high sequence reproducibility (Additional file 5: Table S3). Across clinical isolate replicates the number of SNPs identified was similar and the overlap between variant calling algorithms was high (>90%; Additional file 5: Table S3). This observation was supported by the Ion PGM data but, due to uneven coverage, at least 120 SNPs fewer were identified when compared to matching MiSeq samples. Within platforms and calling algorithms there was variation between replicates in the indels detected, but there was high overlap between algorithms (>90%; Additional file 5: Table S3). Compared to SNPs the breakpoints for these variants are more difficult to characterise from alignments.

For the MiSeq platform data we assessed the number of SNP genotypes called across a range of coverage depths of the reference and alternative alleles (total depth 5- to 20-fold; major allelic frequency >0.5 and >0.7). The number of SNPs decreased pseudo-linearly with decreasing minimum read depth for H37Rv (87 to 67 SNPs; Additional file 6: Figure S3) and the ten clinical isolates (2290 to 2097 SNPs; Additional file 7:

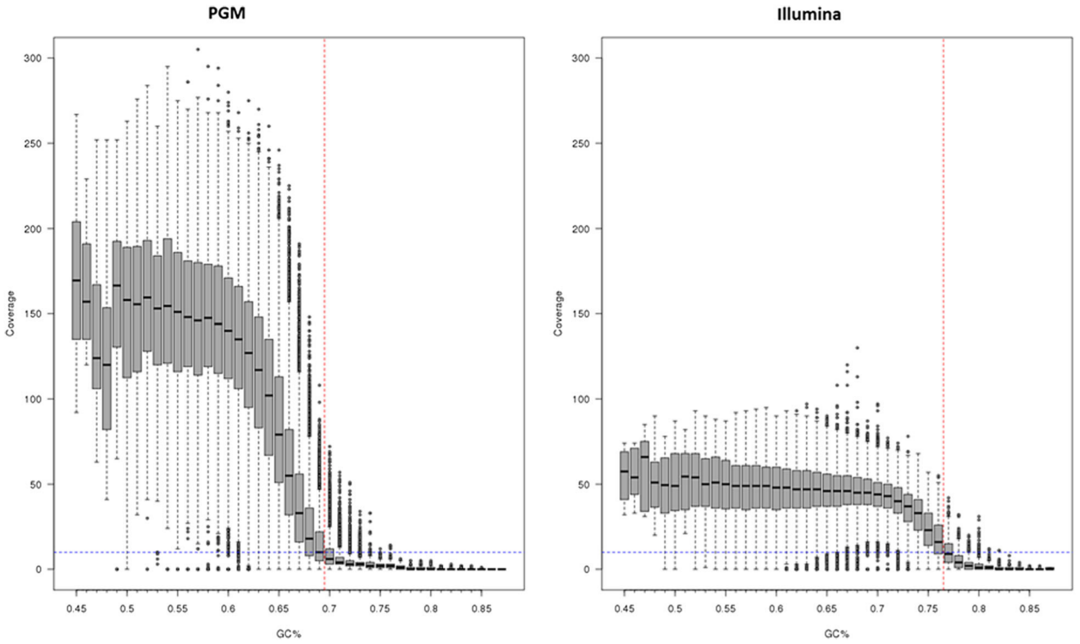


Fig. 1 The dependence of coverage on GC content. The coverage across regions of the genome with differing GC content compared using two different sequencing technologies; the Ion PGM and the Illumina MiSeq. The *dashed blue line* represents the cutoff used when calling variants. Any position which had a coverage <10 was marked as missing. The *dashed red line* shows at which GC% the median coverage across the window falls below the cutoff

Figure S4). In general, differences in the number of SNPs between the *Samtools* and *GATK* algorithms decreased as the depth of coverage and allelic frequency thresholds increased. For H37Rv, read depths in excess of 20-fold had no impact on variants detected. Across clinical isolates, the highest possible stringency tested consisted of using a minimum coverage of 20 and an allelic frequency of 0.7 and led to near identical numbers of total SNPs called by the two variant calling algorithms (*Samtools* 1943, *GATK* 1990, either 2097, both 1840 SNPs; Additional file 7: Figure S4). Much of the discordance in the number of SNPs within replicate groups is due to differences in coverage leading to some polymorphisms not passing quality control filters. Using SNPs for which all replicates have non-missing genotypes, all replicates had identical numbers of SNPs except POR3C, which differed by two SNPs between POR3A and POR3B. Overall, the analyses indicated no major differences in SNPs detected between the two calling algorithms, and this supported the use of consensus variants for downstream analysis. For example, the set of common SNP variants were used to cluster all samples within a phylogenetic tree using *FastTree* v2.1.7 [26] (Additional file 8: Figure S5). Perfect clustering was observed amongst isolates and their replicates. At a finer resolution, we analysed the SNP differences between the replicates, and none were identified.

Calling in silico resistance phenotypes

When the MiSeq raw sequence data were subjected to analysis using *TBProfiler*, agreement with phenotypic susceptibility testing was high (95.3%, 82/86; Table 1). Discrepant results were recorded for PZA (×2) and PAS (×2) where phenotypically resistant isolates not identified by *TBProfiler* were found to have novel mutations in known candidate genes (Additional file 9: Table S4). The novel polymorphisms included a deletion in *pncA* of 20 bp (nucleotides 437–449) and a nucleotide insertion (GG) between codons 130 and 131. PAS-resistant isolates had a *folC* S98G mutation and a *thyX G*-4A, *thyX I161T*, *dfrA-thyA* deletion. Phenotypic testing of kanamycin drug susceptibility was not performed, but mutations associated with its resistance were detected in all eight isolates (Table 1; Additional file 9: Table S4). All mutations were confirmed using independent Sanger capillary sequencing and/or the line probe assays Genotype MTBDR*plus* and Genotype MTBDR*sl* (Hain).

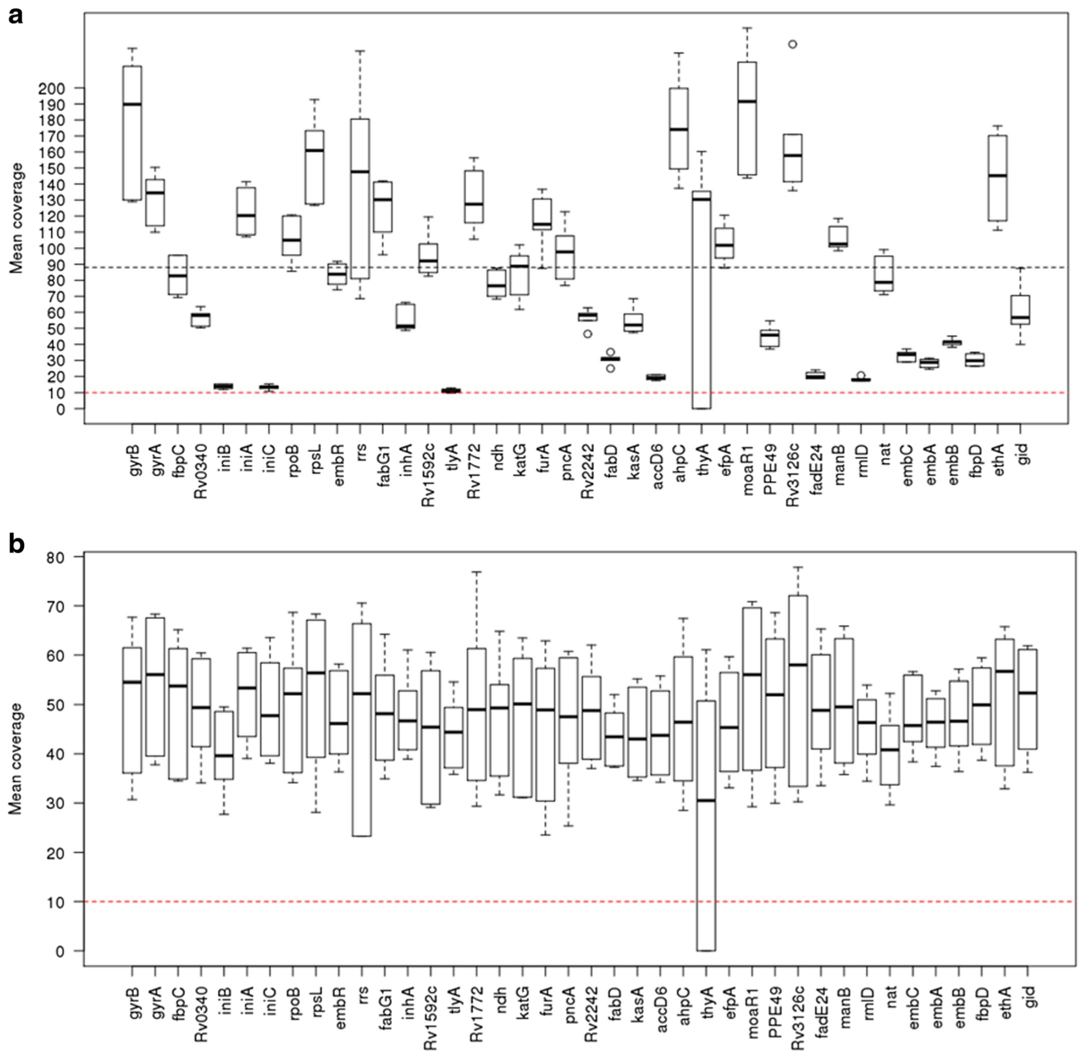


Fig. 2 Coverage across drug-resistance genes. The coverage across the drug-resistance genes in POR1, 2 and 6 samples sequenced using both the **a** Ion PGM and **b** Illumina MiSeq. The *dashed red line* represents the cutoff used when calling variants. Any position with less than tenfold coverage was marked as missing. The low coverage in *thyA* is due to a deletion polymorphism

Phenotypic resistance profiles were confirmed and quantified by the qDST method for the MGIT960 system [12, 13].

The *Mykrobe predictor* tool was also applied to in silico call resistance. This approach looks for mutations associated with resistance to first-line drugs (rifampicin, isoniazid, ethambutol) and second-line drugs (streptomycin, ciprofloxacin, ofloxacin, moxifloxacin, amikacin, kanamycin, capreomycin). Of the 72 resistance calls made, 19 (26.4%) were incorrectly called “susceptible”. False negative calls were made for isoniazid (×1), ethambutol (×2), streptomycin (×4), amikacin (×4), and capreomycin (×3). Additionally there was a disagreement between *TBprofiler* and *Mykrobe predictor* with four samples for kanamycin, the latter program calling them as “susceptible” (Table 1).

For Ion PGM, when predicting individual drug-resistance profiles in the three isolates, in one isolate the *gyrA* D94A mutation associated with fluoroquinolone resistance could not be detected due to lack of coverage

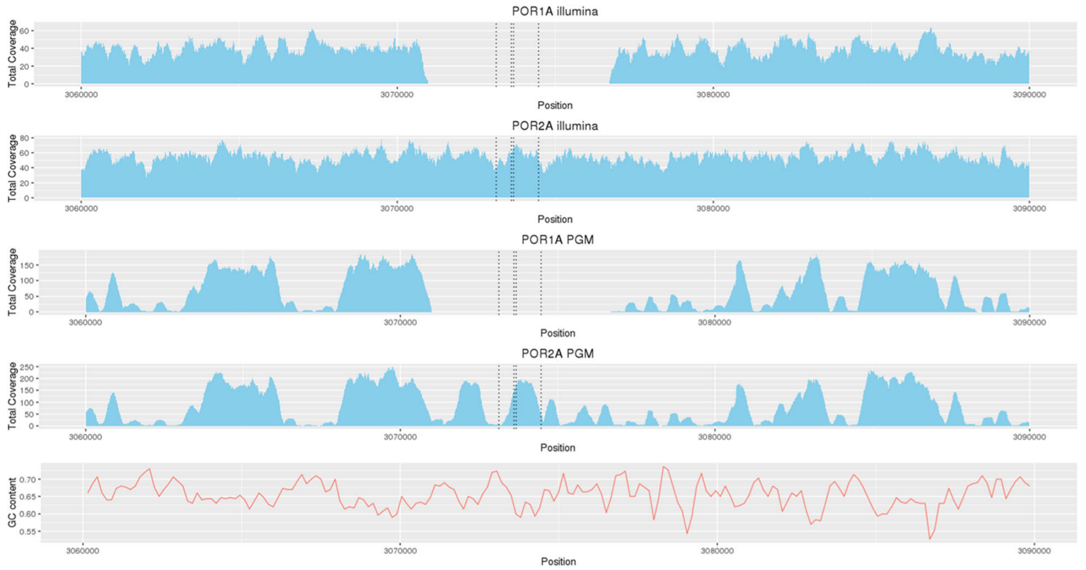


Fig. 3 Lack of genomic coverage in *dfrA-thyA* genes reveals deletions in the POR1A XDR isolate with PAS resistance. Uneven Ion PGM sequence coverage is due to high GC content

(Additional file 5: Table S3). However, the mutation was recovered if the coverage threshold was relaxed from ten- to fourfold.

Discussion

Advances in next-generation sequencing technology have expanded opportunities for genome analysis in the clinical laboratory. Determining resistance to anti-TB drugs by whole genome sequencing has been demonstrated as feasible and is being implemented in some specialist centres [6]. For acceptance as a diagnostic tool to guide treatment of drug-resistant TB the sequencing platforms and analytical tools employed must be robust and reliable. Here we have investigated the performance of two commercial 'bench-top' next generation sequencing platforms and attempted to assess the robustness of a bioinformatics analysis pipeline with respect to variant calling, across sequencing replicates.

The MiSeq and Ion PGM both proved satisfactory for determining drug-resistance profiles. Compared to Ion PGM, MiSeq sequence coverage was more uniform and was better represented in regions with high GC content. However, we did not investigate the impact of the different library preparation methods used (mechanical (MiSeq) and enzymatic (Ion PGM) processing). Sample quality and the mode or preparation have been shown to influence the depth of coverage in high GC regions [27], and further work is required to investigate this. The Ion PGM platform has previously been used to characterise mutations in XDR-TB strains [6], but the minimum read depth used to call alleles (fourfold) were less stringent than the tenfold coverage threshold adopted here.

Samtools and *GATK* when used to process the raw sequence data produced diverse outputs but filtering based on coverage and allelic frequency led to almost complete agreement on resistance causing SNPs. There was, however, lower concordance between the final sets of indels. As previously reported, the false discovery rate for *Samtools* is higher than for *GATK* and rises as coverage increases [28]. A common strategy is to undertake dual analysis and consider the intersection of the *Samtools* and *GATK* derived SNPs but select only the *GATK* indels [16]. The high reproducibility of sequence data from replicate samples is reassuring as it affirms the validity of next-generation sequencing as a tool for investigating transmission events.

Of the two rapid tools examined, the *TBProfiler* gave 100% concordance with phenotypic DST results for INH, RIF, STR, ETB, ETH and the fluoroquinolones. Of the nine PZA-resistant isolates, known resistance SNPs were reported for seven isolates with an insertion and deletion observed for the remaining two. Possible novel resistance mutations were also observed for both the PAS-resistant isolates. The *Mykrobe predictor* detected resistance for nine drugs, of which eight had DST results. Concordance was 100% for RIF, OFX and MOX, but resistance was missed for one or more isolates for the remaining five drugs. Misclassification of resistance of amikacin and capreomycin as susceptible has significant clinical implications as patients may be assigned treatment that is not effective for XDR-TB.

The identification of a PAS resistance-related *dfrA-thyA* double deletion in an XDR-TB sample highlights the need to look at non-SNP variants. Significantly, the laboratory platform being used may impact the detection of putative drug resistance. This is critical in XDR-TB and resistance beyond XDR-TB where use of drugs like PAS may make the difference in providing a life-saving effective regimen of at least five drugs [29]. Large deletions and other structural variants may be detected by applying a combination of complementary approaches (pair-end, split-read and depth of coverage) followed by a validation process involving de novo assembly of bordering reads and re-alignment to the reference genome [10, 16, 24]. However, high genome-wide sequence coverage is necessary to perform such analyses.

As expected the genotypic profiling was concordant with the phenotypic determination of drug-resistance levels confirming the reliability and robustness of the selected genes and mutations as predictors of resistance for almost all drugs tested; with discrepancies still being noticed for PZA and PAS due to lack of enough information on their mechanism of action [12, 30]. Surprisingly, no discrepancies were found for EMB, a drug known to have low correlation between the *emb* genes and phenotypic resistance [12].

Conclusions

Sequencing platforms are becoming more accessible and economical. Our work suggests that they are capable of delivering high quality data regarding resistance to anti-TB drugs but do not all perform to the same standard and quality monitoring is advisable. Further studies are needed to evaluate these analytical tools, which as yet do not have regulatory approval for clinical use. It is expected that drug-resistance profiling using next-generation sequencing will gain accuracy and reliability with the gathering of improved knowledge of the drug-target genes and resistance-causing mutations, including for the new drugs recently approved for the treatment of MDR- and XDR-TB [29, 31]. Ultimately, drug resistance profiling using next-generation sequencing offers rapid assessment of resistance-associated mutations, thus accelerating access to effective treatment.

Additional files

Additional file 1: Figure S1. Bioinformatics pipeline. (TIFF 81 kb)

Additional file 2: Table S1. Sanger sequencing primers for genomic variant confirmation. (DOCX 132 kb)

Additional file 3: Table S2. Summary of the sequencing data, coverage and SNPs for each sample. (DOCX 22 kb)

Additional file 4: Figure S2. a Mean coverage for all samples for each drug resistance gene. Deletion of *dfrA-thyA* is evident by the zero coverage outliers in POR1. **b** Mean coverage across drug-resistance genes. (TIFF 273 kb)

Additional file 5: Table S3. Replicate variation across extraction and calling algorithms, and phenotypic profiles. (DOCX 20 kb)

Additional file 6: Figure S3. The changes in the number of SNPs characterised across algorithms for H37Rv. (TIFF 85 kb)

Additional file 7: Figure S4. The changes in the number of SNPs characterised across algorithms for the ten clinical isolates. (TIFF 92 kb)

Additional file 8: Figure S5. Phylogenetic tree of all the MiSeq sequenced samples. (TIFF 76 kb)

Additional file 9: Table S4. Mutations that potentially explain drug resistance in the samples. (DOCX 19 kb)

Acknowledgements
We wish to thank ThermoFisher Scientific for performing the Ion Torrent PGM™ sequencing. Sequence analysis was performed on the MRC UK funded eMedlab computing resource.

Funding
This work was supported by the UK National Measurement System and the European Metrology Research Programme (EMRP) joint research project [HLT08] "Infect-Met", which is jointly funded by the EMRP participating countries within EURAMET and the European Union. JP is the recipient of a BBSRC PhD studentship. TGC is funded by the Medical Research Council UK (grant number MR/K000551/1, MR/M01360X/1, MR/N010469/1). SC is funded by the Medical Research Council UK (grant number MR/M01360X/1). DM is supported by Fundação para a Ciência e a Tecnologia, Portugal (grant number SFRH/BPD/100688/2014). DM, JR and MV are thankful for the support of grant GHTMUID/Multi/04413/2013 from Fundação para a Ciência e a Tecnologia, Portugal. These funding bodies did not have a role in the design of the study and collection, analysis and interpretation of data and in writing the manuscript.

Availability of data and materials
All raw sequencing data for both MiSeq and Ion PGM can be accessed from the European Nucleotide Archive (ENA) Sequence Read Archive (project accession PRJEB15555). Pipeline scripts are available from http://pathogenseq.lshtm.ac.uk.

Authors' contributions
JP, DMO, MV, JFH and TGC conceived and designed the study; DMO, ASW, DM and JR performed laboratory experiments and curation of meta data for sequencing; DMO, DM, JR, ASW, JO'G, SC, MV and JFH contributed biological samples, sequencing or phenotypic data; JP performed the statistical analysis under the guidance of SC and TGC; DMO and JFH led the sequencing efforts; JP, DMO, DM, KD, RM, MV, JFH and TGC wrote/drafted and finalised the manuscript with contributions from all other authors. The final manuscript was read and approved by all authors.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Author details
[1]Department of Pathogen Molecular Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, WC1E 7HT London, UK. [2]Molecular Biology, LGC Ltd, Queens Road, Teddington, Middlesex TW11 0LY, UK. [3]Unidade de Microbiologia Médica, Global Health and Tropical Medicine, GHTM, Instituto de Higiene e Medicina Tropical, IHMT, Universidade NOVA de Lisboa, UNL, Lisbon, Portugal. [4]Norwich Medical School, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK. [5]Division of Pulmonary Medicine and UCT Lung Institute, Lung Infection and Immunity Unit, University of Cape Town, Groote Schuur Hospital, Observatory, 7925, Cape Town, South Africa. [6]School of

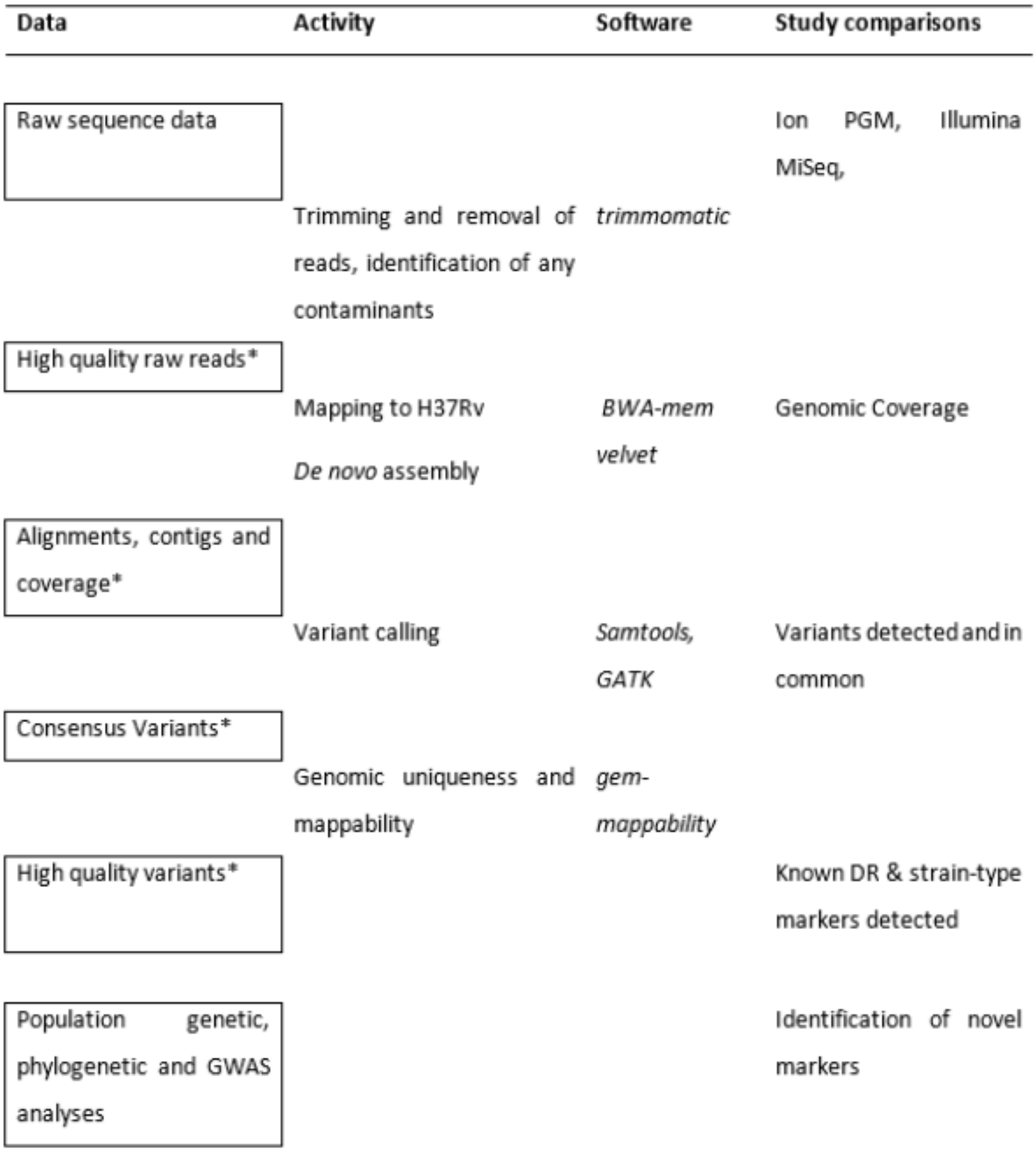
Biosciences & Medicine, Faculty of Health & Medical Science, University of Surrey, Guildford GU2 7XH, UK. [7]Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, WC1E 7HT London, UK.

Received: 30 August 2016 Accepted: 30 November 2016
Published online: 22 December 2016

References

1. World Health Organization. Global Tuberculosis Report 2015. Geneva: World Health Organization; 2015.
2. Zignol M, Dean AS, Falzon D, van Gemert W, Wright A, van Deun A, et al. Twenty years of global surveillance of antituberculosis-drug resistance. N Engl J Med. 2016;375:1081–9.
3. Dheda K, Gumbo T, Gandhi NR, Murray M, Theron G, Udwadia Z, et al. Global control of tuberculosis: from extensively drug-resistant to untreatable tuberculosis. Lancet Respir Med. 2014;2:321–38.
4. Pietersen E, Peter J, Streicher E, Sirgel F, Rockwood N, Mastrapa B, et al. High frequency of resistance, lack of clinical benefit, and poor outcomes in capreomycin treated South African patients with extensively drug-resistant tuberculosis. PLoS One. 2015;10:e0123655.
5. Coll F, McNerney R, Preston M, Guerra-Assunção JA, Warry A, Hill-Cawthorn G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med. 2015;5:51.
6. Witney AA, Gould KA, Arnold A, Coleman D, Delgado R, Dhillon J, et al. Clinical application of whole-genome sequencing to inform treatment for multidrug-resistant tuberculosis cases. J Clin Microbiol. 2015;53:1473–83.
7. Zhang H, Li D, Zhao L, Fleming J, Lin N, Wang T, et al. Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance. Nat Genet. 2013;45:1255–60.
8. Clark TG, Mallard K, Coll F, Preston M, Assefa S, Harris D, et al. Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing. PLoS One. 2013;8:e83012.
9. Farhat MR, Sultana R, Iartchouk O, Bozeman S, Galagan J, Sisk P, et al. Genetic determinants of drug resistance in mycobacterium tuberculosis and their diagnostic value. Am J Respir Crit Care Med. 2016. doi:10.1164/rccm.201510-2091OC.
10. Phelan J, Coll F, McNerney R, Ascher DB, DE Pires V, Furnham N, et al. Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. BMC Med. 2016;14:31.
11. Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZM, et al. Rapid Whole Genome Sequencing of M. tuberculosis directly from clinical samples. J Clin Microbiol. 2015. doi:10.1128/JCM.00486-15.
12. Cambau E, Viveiros M, Machado D, Raskine L, Ritter C, Tortoli E, et al. Revisiting susceptibility testing in MDR-TB by a standardized quantitative phenotypic assessment in a European multicentre study. J Antimicrob Chemother. 2015;70:686–96.
13. Springer B, Lucke K, Calligaris-Maibach R, Ritter C, Bottger EC. Quantitative drug susceptibility testing of Mycobacterium tuberculosis by use of MGIT 960 and EpiCenter instrumentation. J Clin Microbiol. 2009;47:1773–80.
14. Larsen MH, Biermann K, Tandberg S, Hsu T, Jacobs WR, Larsen MH. Genetic manipulation of *Mycobacterium tuberculosis*. Curr Protoc Microbiol. 2007;8:10A.2.1.
15. Benavente ED, Coll F, Furnham N, McNerney R, Glynn JR, Campino S, et al. PhyTB: Phylogenetic tree visualisation and sample positioning for M tuberculosis. BMC Bioinformatics. 2015;16:155.
16. Coll F, Preston M, Guerra-Assunção JA, Hill-Cawthorn G, Harris D, Perdigão J, et al. PolyTB: a genomic variation map for Mycobacterium tuberculosis. Tuberculosis (Edinb). 2014;94:346–54.
17. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30:2843–51.
18. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:2987–93.
19. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
20. Preston MD, Campino S, Assefa SA, Echeverry DF, Ocholla H, Amambua-Ngwa A, et al. A barcode of organellar genome polymorphisms identifies the geographic origin of Plasmodium falciparum strains. Nat Commun. 2014. doi:10.1038/ncomms5052.
21. Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, et al. Fast computation and applications of genome mappability. PLoS One. 2012;7:e30377.
22. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, et al. Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. Nat Commun. 2015;6:10063.
23. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539.
24. Phelan JE, Coll F, Bergval I, Anthony RM, Warren R, Sampson SL, et al. Recombination in pe/ppe genes contributes to genetic variation in Mycobacterium tuberculosis lineages. BMC Genomics. 2015;17:151.
25. Moradigaravand D, Grandjean L, Martinez E, Li H, Zheng J, Coronel J, et al. dfrA thyA Double deletion in para-aminosalicylic acid-resistant Mycobacterium tuberculosis Beijing strains. Antimicrob Agents Chemother. 2016;60:3864–7.
26. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009;26:1641–50.
27. Tyler AD, Christianson S, Knox NC, Mabon P, Wolfe J, Van Domselaar G, et al. Comparison of sample preparation methods used for the next-generation sequencing of Mycobacterium tuberculosis. PLoS One. 2016;11:e0148676.
28. Liu X, Han S, Wang Z, Gelernter J, Yang B-Z, Nielsen R, et al. Variant callers for next-generation sequencing data: a comparison study. PLoS One. 2013;8:e75619.
29. WHO. WHO treatment guidelines for drug-resistant tuberculosis. 2016. http://www.who.int/tb/areas-of-work/drug-resistant-tb/treatment/resources/en/. Accessed 10 Oct 2016.
30. Domínguez J, Boettger EC, Cirillo D, Cobelens F, Eisenach KD, Gagneux S, et al. Clinical implications of molecular drug resistance testing for Mycobacterium tuberculosis: a TBNET/RESIST-TB consensus statement. Int J Tuberc Lung Dis. 2016;20:24–42.
31. Papaventsis D, Casali N, Kontsevaya I, Drobniewski F, Cirillo DM, Nikolayevskyy V. Whole genome sequencing of M. tuberculosis for detection of drug resistance: a systematic review. Clin Microbiol Infect. 2016. doi:10.1016/j.cmi.2016.09.008.

Additional File 1: Figure S1
Bioinformatics pipeline



Data	Activity	Software	Study comparisons
Raw sequence data			Ion PGM, Illumina MiSeq,
	Trimming and removal of reads, identification of any contaminants	*trimmomatic*	
High quality raw reads*			
	Mapping to H37Rv *De novo* assembly	*BWA-mem* *velvet*	Genomic Coverage
Alignments, contigs and coverage*			
	Variant calling	*Samtools, GATK*	Variants detected and in common
Consensus Variants*			
	Genomic uniqueness and mappability	*gem-mappability*	
High quality variants*			Known DR & strain-type markers detected
Population genetic, phylogenetic and GWAS analyses			Identification of novel markers

* Drug resistance (DR) mutations and strain-types rapidly identified from these data

Additional File 2: Table S1
Sanger sequencing primers for genomic variant confirmation

Gene	Primer	Primer sequence (5’-3’)	Annealing (^{0}C)	Length (bp)	Ref.
inhA	*inhA-1*	CCT CGC TGC CCA GAA AGG GA	64	248	A
	inhA-2	ATC CCC CGG TTT CCT CCG GT			
	inhA-3	AGG TCG CCG GGG TGG TCA GC	60	517	
	inhA-4	AGC GCC TTG GCC ATC GAA GCA			
	inhA-3F	CCA CAT CTC GGC GTA TTC G		501	B
	inhA-5R	TTC CGG TCC GCC GAA CGA CAG			
katG	*P4_Fw*	CGG ACC ATA ACG GCT TCC TG	62	563	C
	P4_Rv	TTG TCC AAG CTG GCG TTG TC			
	P5_Fw	CGA CAA CGC CAG CTT GGA C		518	
	P5_Rv	CGG TTC CGG TGC CAT ACG			
	P6_Fw	AGC TCG TAT GGC ACC GGA AC		619	
	P6_Rv	TGA CCT CCC ACC CGA CTT GT			
	P7_Fw	ACA AGT CGG GTG GGA GGT C		574	
	P7_Rv	CTG CCG GTC CAC TTC ACC TT			
	P8_Fw	GGG ACC TAC CAG GGC AAG GA		629	
	P8_Rv	CCG GGA GTC AGC AAG TCA CC			
tlyA	*tlyAs*	GCA TCG CAC GTC GTC TTT	55	947	D
	tlyAas	GGT CTC GGT GGC TTC GTC			
eis	*eisF1*	GCC ATG GGA CCG GTA CTT GC	56	601	E
	eisR1	GTA GAT GCC GCC CTC GCT AG			
gidB	*gidB_Fw*	CGA GAG CGG AGA ATG TTT CA	62	793	F
	gidB_Rv	CTG GCC CGA CCT TAC GAG			

pncA	*pncA_promP1*	GCT GGT CAT GTT CGC GAT CG	55	214	G
	pncA_promP2	TCG GCC AGG TAG TCG CTG AT			
	pncA_Fw	AGT CGC CCG AAC GTA TGG TG	62	615	H
	pncA_Rv	CAA CAG TTC ATC CCG GTT CG			
gyrA	*gyrA_Fw*	ATC GCC GGG TGC TCT ATG	62	321	F
	gyrA_Rv	GGC CGT CGT AGT TAG GGA TG			
ethA	*ethA1*	ATC ATC GTC GTC TGA CTA TGG	55	667	A
	ethA5	ACT ACA ACC CCT GGG ACC			
	ethA4	CCT CGA CCT TCC CGT GA	64	692	
	ethA9	CCT CGA GTA CGT CAA GAG CAC			
	ethA8	GGT GGA ACC GGA TAT GCC TG	68	342	
	ethA10	CGT TGA CGG CCT CGA CAT TAC			
embB	*embB-F2*	AAC CTG CGC CCG CAG ATT GTC	62	526	I
	embB-R2	GGT CTG GCA GGC GCA TCC			
	embBR2_Fw	CTG GCG CTG ATG ACC CAT	62	588	*
	embBR2_Rv	GGT GGG CAG GAT GAG GTA G			
***embC-embA* IRG**	*embC-embA_Fw*	GGT TGA CGC CTT ACT ACC C	62	535	J
	embC-embA_Rv	CCA CGA CGA CCG TGT CC			
rpsL	*rpsL_Fw*	GGC CGA CAA ACA GAA CGT	64	504	K
	rpsL_Rv	GTT CAC CAA CTG GGT GAC			
folC	*folCP1*-Fw	CGC TGC AAT GAA TTC GAC GA	62	668	*
	folCP1-Rv	TGA TGA TGC CCG CCT TCT C			
thyX	*thyX*prom_Fw	TGG ATG GAA AAC CTT GCG G	62	558	*
	*thyX*prom_Rv	TCG GTC TTG GCG ATC AGT T			
	thyX-F2	CTA CTC GCA GCT CTC CCA G	62	510	*
	thyX-R2	TAC CTG GCG CTT TAT CCC G			

[A] Morlock G, Metchock B, Sikes D, Crawford J, Cooksey R. ethA, inhA, and katG loci of ethionamide-resistant clinical *Mycobacterium tuberculosis* isolates. Antimicrob Agents Chemother 2003; 47: 3799-805.

[B] Leung E, Ho P, Yuen K, Woo W, Lam T, Kao R, Seto W, Yam W. Molecular characterization of isoniazid resistance in *Mycobacterium tuberculosis*: identification of a novel mutation in inhA. Antimicrob Agents Chemother 2006; 50: 1075-8.

[C] Machado D, Perdigão J, Ramos J, Couto I, Portugal I, Ritter C, Boettger E, Viveiros M. High-level resistance to isoniazid and ethionamide in multidrug-resistant *Mycobacterium tuberculosis* of the Lisboa family is associated with inhA double mutations. J Antimicrob Chemother. 2013; 68: 1728-32.

[D] Feuerriegel S, Cox H, Zarkua N, Karimovich H, Braker K, Rüsch-Gerdes S, Niemann S. Sequence analyses of just four genes to detect extensively drug-resistant *Mycobacterium tuberculosis* strains in multidrug-resistant tuberculosis patients undergoing treatment. Antimicrob Agents Chemother. 2009; 53: 3353-6.

[E] Perdigão J, Macedo R, Silva C, Machado D, Couto I, Viveiros M, Jordão L, Portugal I. From multidrug-resistant to extensively drug-resistant tuberculosis in Lisbon, Portugal: the stepwise mode of resistance acquisition. J Antimicrob Chemother. 2013; 68: 27-33

[F] Machado, D. (2014). The dynamics of drug resistance in *Mycobacterium tuberculosis*: exploring the biological basis of multi- and extensively drug resistant tuberculosis (MDR/XDRTB) as a route for alternative therapeutic strategies. PhD thesis. Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa. Lisboa, Portugal.

[G] Scorpio A, Lindholm-Levy P, Heifets L, Gilman R, Siddiqi S, Cynamom M, Zhang Y. Characterization of pncA mutations in pirazinamide-resistant *Mycobacterium tuberculosis*. Antimicrob Agents Chemother. 1997; 41: 540-3.

[H] Louw G, Warren R, Donald P, Murray M, Bosman M, Van Helden P, Young D, Victor T. Frequency and implications of pyrazinamide resistance in managing previously treated tuberculosis patients. Int J Tuberc Lung Dis. 2006; 10: 802-7.

[I] Starks A, Gumusboga A. Plikaytis B, Shinnick T, Posey J. Mutations at embB306 are an important molecular indicator of ethambutol resistance in *Mycobacterium tuberculosis*. Antimicrob Agents Chemother. 2009; 53: 1061-66.

[J] Cui Z, Li Y, Cheng S, Yang H, Lu Junmei, Hu Z, Ge B. Mutations in the embC-embA intergenic region contribute to *Mycobacterium tuberculosis* resistance to ethambutol. Antimicrob Agents Chemother. 2014; 58: 6837-43.

[K] Sreevatsan S, Pan X, Stockbauer K, Williams D, Kreiswirth B, Musser J. Characterization of rpsL and rrs mutations in streptomycin-resistant *Mycobacterium tuberculosis* isolates from diverse geographic localities. Antimicrob Agents Chemother. 1996; 40: 1024-6.

* This work

Additional File 3: Table S2
Summary of the sequencing data, coverage and SNPs for each sample

Sequencing platform	Sample	No. reads	Median read length	Proportion coverage > 10-fold	Median coverage	Total SNPs
MiSeq	POR1A	874721	222	0.95	40	766
MiSeq	POR1B	1280618	221	0.96	55	766
MiSeq	POR1C	1068336	221	0.96	48	766
Ion PGM	POR1A	1015193	335	0.73	48	512
Ion PGM	POR1B	1124215	339	0.67	52	512
MiSeq	POR2A	1167341	224	0.97	53	854
MiSeq	POR2B	871084	223	0.97	38	854
MiSeq	POR2C	817606	224	0.97	36	854
Ion PGM	POR2A	929733	213	0.73	28	594
Ion PGM	POR2C	966514	326	0.74	46	594
MiSeq	POR3A	1217694	224	0.96	55	771
MiSeq	POR3B	1100251	222	0.96	50	771
MiSeq	POR3C	413660	215	0.93	18	773
MiSeq	POR4A	1055194	218	0.96	47	795
MiSeq	POR4B	1100448	224	0.96	50	795
MiSeq	POR4C	1071269	225	0.96	49	795
MiSeq	POR5A*	988848	224	0.96	45	758
MiSeq	POR5B	1111052	224	0.96	51	758
MiSeq	POR5C	1113854	223	0.96	50	758
MiSeq	POR6A*	2269310	180	0.97	70	767
MiSeq	POR6B	1201932	222	0.96	53	767
MiSeq	POR6C	774063	222	0.96	34	767
Ion PGM	POR6B	1049314	338	0.72	44	510
Ion PGM	POR6C	904304	325	0.73	42	510
MiSeq	POR7A*	2423026	179	0.97	70	801
MiSeq	POR7B	1129806	222	0.96	51	801
MiSeq	POR7C	2638858	155	0.97	65	801
MiSeq	POR8A	2851160	172	0.97	79	770
MiSeq	POR8B	1028634	225	0.96	49	770
MiSeq	POR8C	801687	222	0.96	36	770
MiSeq	POR9A	2091394	180	0.97	61	796
MiSeq	POR9B	1145983	225	0.96	53	796
MiSeq	POR9C	1128251	223	0.96	51	796
MiSeq	POR10A	1074170	217	0.97	48	902
MiSeq	POR10B	1224053	223	0.97	54	902
MiSeq	POR10C	894289	223	0.97	39	902
MiSeq	H37Rv*	2652971	156	0.99	56	62

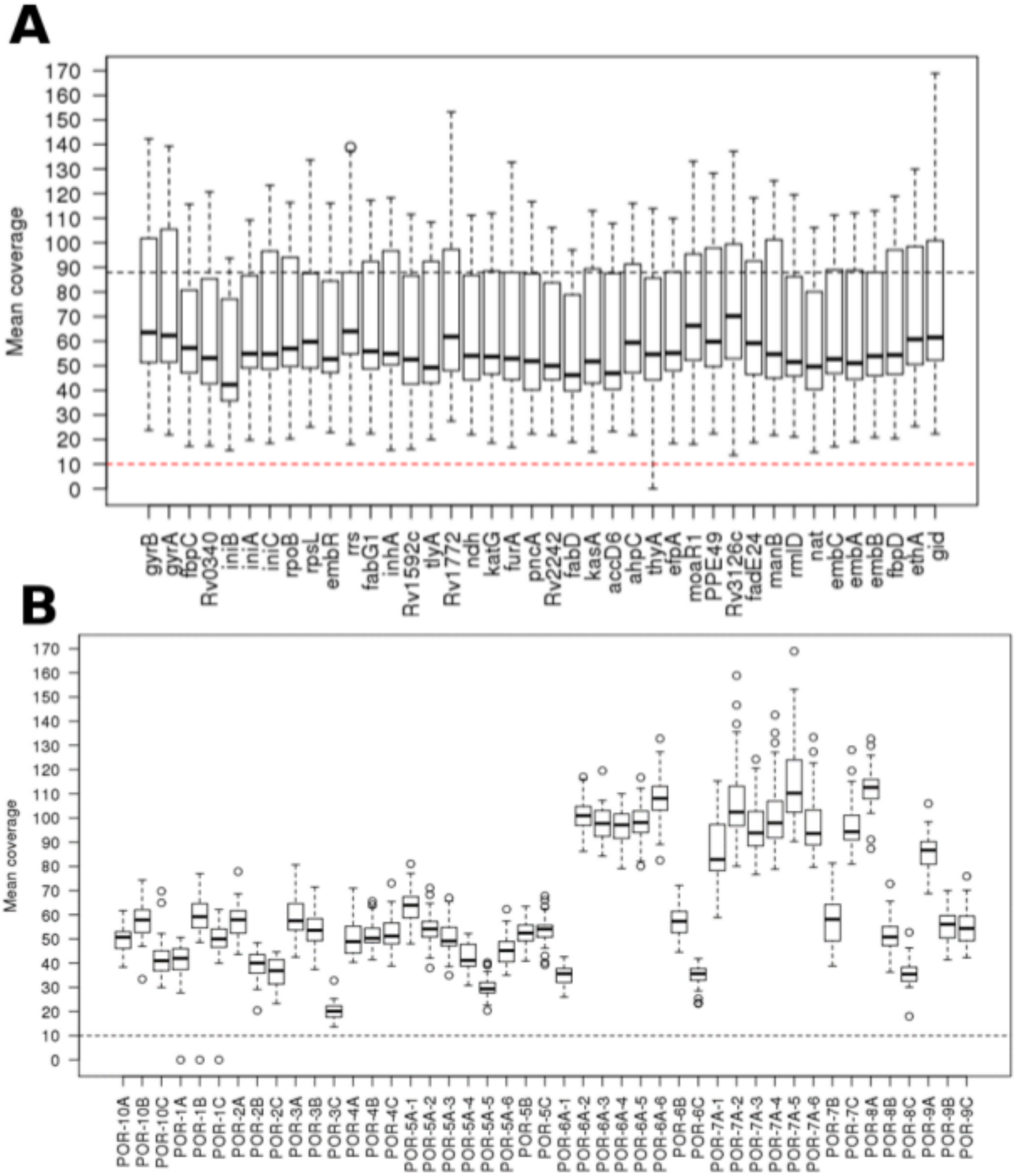
* 6 technical replicates for each, and average statistics presented; A-C refers to extraction replicates of the same samples

Additional File 4: Figure S2

(a) Mean coverage for all samples for each drug resistance gene

Deletion of *dfrA-thyA* is evident by the zero coverage outliers in POR1

(b) Mean coverage across drug resistance genes



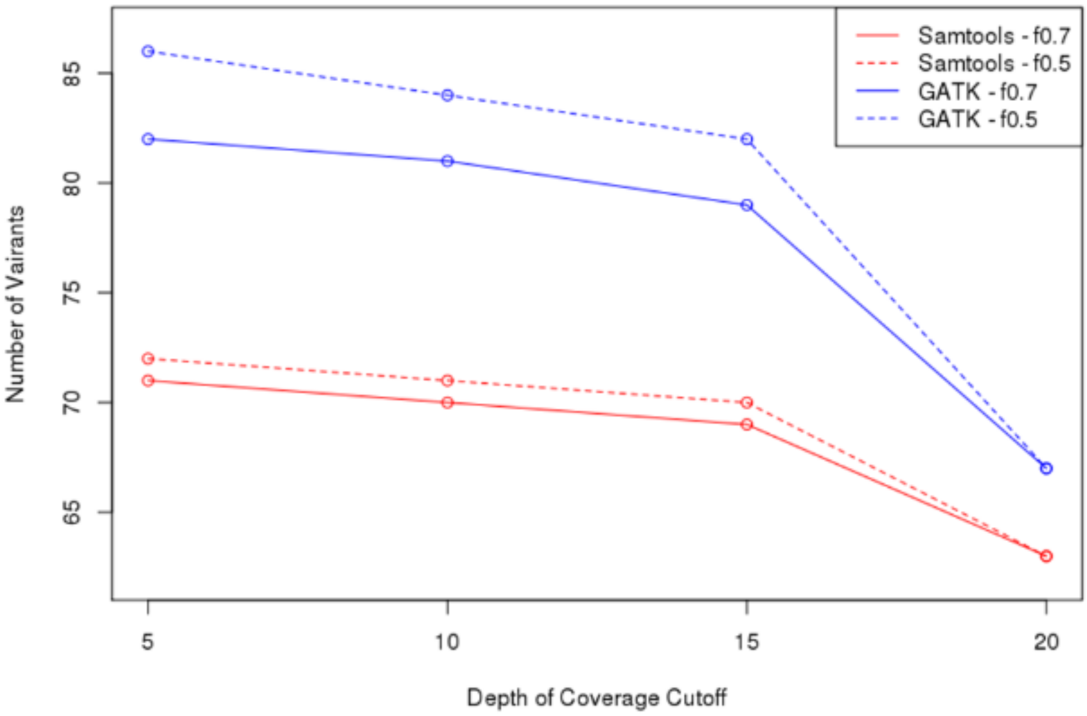
Additional File 5: Table S3
Replicate variation across extraction and calling algorithms, and phenotypic profiles

Sequencing platform	**Comparison (no. rep, M/XDR-TB)**	**GATK SNPs***	***Samtools* SNPs***	***Overlap SNPs %***	***GATK* Indels Min. / total (overlap)**	***Samtools* Indels Min. / total (overlap)**	***Overlap Indels %***	**Inferred MDR/XDR-TB**	**Inferred drug resistance*** INH, RIF, ETH +**
	Technical								
MiSeq	POR5A (6,M)	783	753	96.2	85/98 (0.87)	66/104 (0.63)	94.2	MDR-TB	ETB, PZA, STR
MiSeq	POR6A (6,X)	846	815	92.2	87/103 (0.84)	65/127 (0.51)	81.1	XDR-TB	ETB, PZA, STR, FLQ, AMK, CAP, KAN
MiSeq	POR7A (6,X)	858	839	97.1	90/102 (0.88)	72/133 (0.54)	76.7	XDR-TB	ETB, PZA, STR, FLQ, AMK, CAP, KAN
MiSeq	H37Rv (6,S)	81	70	84.1	22/27 (0.81)	16/40 (0.40)	67.5	Susc.	
	Extraction								
MiSeq	POR1 (3,X)	788	753	95.6	86/91 (0.95)	72/99 (0.73)	91.9	XDR-TB	ETB, PZA, STR, FLQ, AMK, CAP, KAN
Ion PGM	POR1 (2,X)	618	611	95.7	48/98(0.49)	53/81 (0.65)	34.1	XDR-TB	ETB, PZA, STR, FLQ, AMK, CAP, KAN
MiSeq	POR2 (3,M)	875	846	96.7	99/114 (0.87)	88/115 (0.77)	99.1	MDR-TB	
Ion PGM	POR2 (2,M)	710	706	96.3	23/52(0.44)	39/67 (0.58)	22.2	MDR-TB	
MiSeq	POR3 (3,X)	804	788	97.8	87/98 (0.89)	70/100 (0.70)	98.0	XDR-TB	ETB, STR, FLQ, AMK, CAP, KAN
MiSeq	POR4 (3,X)	805	789	98.0	86/91 (0.95)	69/92 (0.75)	98.9	XDR-TB	ETB, PZA, STR, FLQ, AMK, KAN
MiSeq	POR5 (3,M)	784	754	96.2	90/92 (0.98)	74/94 (0.79)	97.9	MDR-TB	ETB, PZA, STR

MiSeq	POR6 (3,X)	849	827	90.9	87/98 (0.89)	70/97 (0.72)	99.0	XDR-TB	ETB, PZA, STR, FLQ, AMK, CAP, KAN
Ion PGM	POR6 (2,X)	617	612	95.7	33/76(0.43)	46/82 (0.56)	23.8	**MDR-TB**	ETB, PZA, STR, AMK, CAP, KAN
MiSeq	POR7 (3,X)	875	868	93.5	87/102 (0.85)	73/104 (0.70)	98.1	XDR-TB	ETB, PZA, STR, FLQ, AMK, CAP, KAN
MiSeq	POR8 (3,X)	820	791	94.1	84/96 (0.88)	74/95 (0.78)	99.0	XDR-TB	ETB, PZA, STR, FLQ, AMK, CAP, KAN
MiSeq	POR9 (3,X)	820	807	97.2	90/98 (0.92)	77/104 (0.57)	94.2	XDR-TB	ETB, PZA, STR, FLQ, AMK, CAP, KAN
MiSeq	POR10 (3,M)	922	885	95.8	98/108 (0.91)	78/107 (0.73)	99.1	MDR-TB	ETB, STR

* Differences between replicates were only due to low coverage missing genotypes i.e. no differing base calls; ** based on comparing all indels detected by each method; *** based on *TBProfiler;* INH Isoniazid, RIF Rifampicin, STR Streptomycin, ETB Ethambutol, PZA Pyrazinamide, RFB Rifabutin, ETH Ethionamide, AMK Amikacin, CAP Capreomycin, OFX Ofloxacin, MOX Moxifloxacin, PAS Para-aminosalicylic acid, LZ Linezolid, KAN Kanamycin; **bold** - Fluoroquinolone (FLQ) resistance mutation *gyrA* D94A was not found

Additional File 6: Figure S3
The changes in the number of SNPs characterised across algorithms for H37Rv



The total number of SNPs for H37Rv isolates called using different algorithms and depths (≥5, ≥10, ≥14, ≥20) and allelic frequency cut-offs (≥0.5, ≥0.7). With read depth ≥10, the allelic frequency cut-offs had no impact on variants detected

Additional file 7: Figure S4

The changes in the number of SNPs characterised across algorithms for the ten clinical isolates

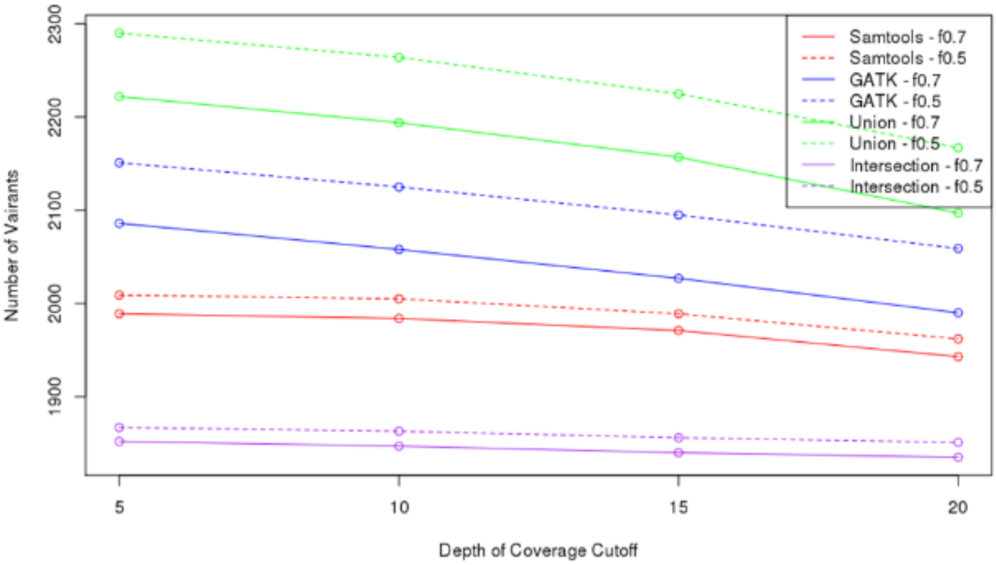
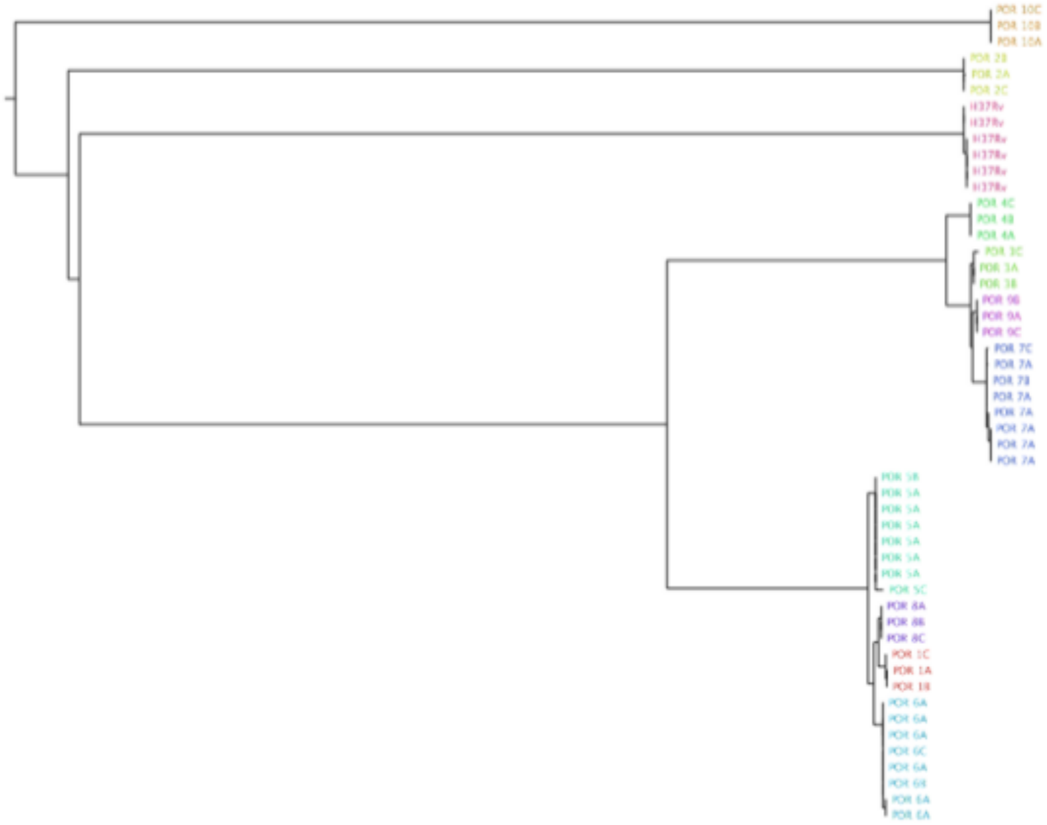


Figure showing the total number of SNPs called using different algorithms and different depth (≥5, ≥10, ≥14, ≥20) and allelic frequency cut-offs (≥0.5, ≥0.7).

Additional File 8: Figure S5
Phylogenetic tree of all the MiSeq sequenced samples



Perfect clustering can be observed across conditions. Each sample is represented by a different colour; replicates of the same patient are shown as the same colour.

Additional File 9: Table S4
Mutations that potentially explain drug resistance in the samples

Sample – M/XDR-TB	**INH***	**RIF***	**STR***	**ETB*,£,**	**PZA***	**ETH*,****	**FLQ*,£**	**AMINO*,£**	**PAS**
POR1 –X	*fabG1_pro* C-15T, *inhA* I194T	*rpoB* S450L	*gidB* A80P	*embA_pro* C-16T, *embB* M306V/M423T	*pncA* V125G	*fabG1_pro* C-15T, *inhA* I194T	*gyrA* D94A	*rrs* A1401G	***thyX* G-4A, *thyX I161T*,** *dfrA-thyA* deletion
POR2-M	*inhA* I21V, *katG* S460N	*rpoB* S450L	-	-	-	*inhA* I21V	-	-	-
POR3- X	*fabG1_pro* C-15T, *inhA* S94A	*rpoB* S450L	*rpsL* K43R	*embA_pro* C-12/11AA, *embB* P397T	**Frameshift mutation *pncA* deletion of nucleotides 437-449**	*fabG1_pro* C-15T, *inhA* S94A	*gyrA* S91P	*tlyA* Ins251TG, *eis_pro* G-10A	-
POR4 –X	*fabG1_pro* C-15T, *inhA* S94A	*rpoB* S450L	*rpsL* K43R	*embB* M306V	*pncA* L120P	*fabG1_pro* C-15T, *inhA* S94A	*gyrA* D94G	*eis_pro* G-10A	-
POR5 -M	*fabG1_pro* C-15T, *inhA* I194T	*rpoB* S450L,	*gidB* A80P	*embB* M306V/M423T	*pncA* V125G	*fabG1_pro* C-15T, *inhA* I194T	-	-	-
POR6 –X	*fabG1_pro* C-15T, *inhA* I194T	*rpoB* S450L	*gidB* A80P	*embA_pro* C-16T, *embB* M306V	*pncA* V125G	*fabG1_pro* C-15T, *inhA* I194T	*gyrA* D94A	*rrs* A1401G	-

POR7 – X	*fabG1_pro* C-15T, *inhA* S94A	*rpoB* S450L,	*rpsL* K43R	*embA_pro* C-12/11AA, *embB* P397T	*pncA* M1T	*fabG1_pro* C-15T, *inhA* S94A	*gyrA* S91P	*tlyA* Ins251TG, *eis_pro* G-10A	-
POR8 – X	*fabG1_pro* C-15T, *inhA* I194T	*rpoB* S450L	*gidB* A80P	*embA_pro* C-16T, *embB* M306V/M423T	*pncA* V125G	*fabG1_pro* C-15T, *inhA* I194T	*gyrA* D94A	*rrs* A1401G	-
POR9 - X	*fabG1_pro* C-15T, *inhA* S94A	*rpoB* S450L	*rpsL* K43R	*embA_pro* C-12/11AA, *embB* P397T	*pncA* M1T	*fabG1_pro* C-15T, *inhA* S94A	*gyrA* S91P	*tlyA* Ins251TG, *eis_pro* G-10A	***folC* S98G**
POR10-M	*fabG1_pro* C-15T, *katG* S315T	*rpoB* S450L	*rpsL* K88R	*embB* S297A/M306I	**GG insertion codons 130 and 131 on *pncA***	*fabG1_pro* C-15T, *ethA* H281P	-	*eis_pro* C-12T	-

All mutations on the positive strand; Confirmed using * Sanger sequencing; ** Genotype MTBDR*plus*; £ Genotype MTBDR*sl*, -- *gyrA* S95T, G668A and *gidB* L16R present, but not resistant related; *rpoB* mutations assigned according to *M. tuberculosis* numbering; potentially novel mutations **bolded;** INH isoniazid, RIF rifampicin, STR Streptomycin, ETB Ethambutol, PZA Pyrazinamide, ETH Ethionamide, FLQ fluoroquinolones (Ofloxacin, Moxifloxacin), AMINO Aminoglycosides (Amikacin, Capreomycin, Kanamycin); PAS Para-aminosalicylic acid

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Jody Phelan
Principal Supervisor	Taane Clark
Thesis Title	A Bioinformatic analysis of *M. tuberculosis* and host genomic data

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?	BMC Medicine		
When was the work published?	March 2016		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I received the raw sequencing data and phenotypic data from our collaborators. I then designed and wrote all scripts to perform the basic data QC, mapping, and variant calling. I optimised the parameters for the GWAS and tabulated the final results in excel. Additionally, I loaded the results in R and produced the figures. After receiving guidance from co-authors and receiving analysis scripts for the protein modelling work, I performed all the protein model QC and ligand docking. I then used scripts provided to me by Davis Ascher to profile the protein stability changes. I wrote the first draft of the manuscript and circulated to co-authors. Upon receiving feedback I revised the manuscript together with Taane Clark and submitted to the journal. I also dealt with subsequent revisions.

Student Signature: ______________________________ **Date:** __________

Supervisor Signature: ______________________________ **Date:** __________

Chapter 3

Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculous drug resistance

Phelan *et al. BMC Medicine* (2016) 14:31
DOI 10.1186/s12916-016-0575-9

BMC Medicine

World TB Day

RESEARCH ARTICLE

Open Access

Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance

Jody Phelan[1], Francesc Coll[1], Ruth McNerney[1,2], David B. Ascher[3], Douglas E. V. Pires[4], Nick Furnham[1], Nele Coeck[5], Grant A. Hill-Cawthorne[6,7], Mridul B. Nair[6], Kim Mallard[1], Andrew Ramsay[8], Susana Campino[1], Martin L. Hibberd[1], Arnab Pain[6], Leen Rigouts[5,9] and Taane G. Clark[1,10,11*]

Abstract

Background: Combating the spread of drug resistant tuberculosis is a global health priority. Whole genome association studies are being applied to identify genetic determinants of resistance to anti-tuberculosis drugs. Protein structure and interaction modelling are used to understand the functional effects of putative mutations and provide insight into the molecular mechanisms leading to resistance.

Methods: To investigate the potential utility of these approaches, we analysed the genomes of 144 *Mycobacterium tuberculosis* clinical isolates from The Special Programme for Research and Training in Tropical Diseases (TDR) collection sourced from 20 countries in four continents. A genome-wide approach was applied to 127 isolates to identify polymorphisms associated with minimum inhibitory concentrations for first-line anti-tuberculosis drugs. In addition, the effect of identified candidate mutations on protein stability and interactions was assessed quantitatively with well-established computational methods.

Results: The analysis revealed that mutations in the genes *rpoB* (rifampicin), *katG* (isoniazid), *inhA*-promoter (isoniazid), *rpsL* (streptomycin) and *embB* (ethambutol) were responsible for the majority of resistance observed. A subset of the mutations identified in *rpoB* and *katG* were predicted to affect protein stability. Further, a strong direct correlation was observed between the minimum inhibitory concentration values and the distance of the mutated residues in the three-dimensional structures of *rpoB* and *katG* to their respective drugs binding sites.

Conclusions: Using the TDR resource, we demonstrate the usefulness of whole genome association and convergent evolution approaches to detect known and potentially novel mutations associated with drug resistance. Further, protein structural modelling could provide a means of predicting the impact of polymorphisms on drug efficacy in the absence of phenotypic data. These approaches could ultimately lead to novel resistance mutations to improve the design of tuberculosis control measures, such as diagnostics, and inform patient management.

Keywords: Tuberculosis, Drug resistance, Genomics, Protein structural modelling, Association study, Convergent evolution

* Correspondence: taane.clark@lshtm.ac.uk
Francesc Coll and Ruth McNerney are joint second authors. Arnab Pain, Leen Rigouts and Taane G Clark are joint senior authors.
[1]Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK
[10]Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK
Full list of author information is available at the end of the article

Background

Tuberculosis, caused by *Mycobacterium tuberculosis* (Mtb), is an important global public health issue (>8.7 million new cases, 1.4 million deaths each year [1]). The *M. tuberculosis* phylogeny consists of four major lineages (L1 - Indo-Oceanic, L2 - East-Asian, L3 - East-African-Indian, L4 - Euro-American), which may vary in their propensity to transmit and cause disease [2]. The Mtb genome (size 4.4 Mb, GC content 65.5 %) is relatively clonal compared to most other bacteria, with no horizontal transfer, and low mutation and recombination rates [3]. Mtb drug resistance is a serious challenge to effective control [1]. Standard first-line anti-TB therapy involves four drugs (rifampicin [RMP], isoniazid [isonicotinic acid hydrazide] [INH], ethambutol [EM]), pyrazinamide [PZA]), with streptomycin (SM) more commonly used when treatment fails. Resistance to at least RMP and INH is denoted as multi drug-resistance (MDR-TB). It has been estimated that ~4 % of new cases are MDR-TB [1], and without effective treatment can remain a source of transmission [4]. Additional resistance to any fluoroquinolone and second-line injectable drug (e.g. amikacin, kanamycin, capreomycin), is denoted as extensively drug resistance (XDR-TB), and such cases have been reported in 100 countries [1].

In routine diagnostic practice susceptibility to anti-tuberculosis drugs is assessed phenotypically by determining the proportion of bacteria that will grow at critical concentrations of the drug [5]. For most anti-tuberculosis drugs, a single concentration is used, but for some drugs two concentrations are used to indicate high and low levels of resistance, where increasing the patient dose may be of clinical benefit. Tests may be performed on solid or liquid media and drug concentrations used may vary according to type of the media and method used. The use of binary reporting (sensitive/resistant) of drug susceptibility, whilst useful for programmic treatment does not inform about the degree of resistance. Minimum inhibitory concentrations (MICs) are determined in some research laboratories where the bacilli are cultured over a range of drug concentrations [6]. Variation in methods and the critical concentrations used creates some disparity between laboratories, particularly for strains where the level of resistance is close to the critical concentration for the drug.

Mtb drug resistance is predominantly conferred by the accumulation of mutations (single nucleotide polymorphisms [SNPs], insertions and deletions [indels]) in genes coding for drug-targets or -converting enzymes [7]. To overcome a loss of fitness that arises during the accumulation of such mutations, putative compensatory mechanisms have been described [8–10]. Many mutations conferring drug resistance have been characterized, especially to first-line treatments [11], and their detection offers a means of rapidly assessing susceptibility to anti tuberculosis drugs to improve patient management [11, 12]. However, with the exception of RMP and INH, current molecular tests for resistance lack sensitivity [7]. RMP is a semisynthetic antibiotic that binds to the RNA polymerase β subunit encoded by *rpoB*, inhibiting transcription. Mutations in *rpoB* can cause resistance to RMP [13]. Mutations occur more frequently in an 81 bp region of the gene termed the RMP resistance determining region [14, 15], and contribute to 96 % of resistance phenotypes (predominantly high level), with S450L (*M. tuberculosis* nomenclature) being the most prevalent mutation [16, 17]. It should be noted however that not all mutations result in the same degree of resistance. For example, substitution of histidine with non-polar leucine (H445L) has a much reduced impact compared to the negatively charged aspartate (H445D) (MIC ~2 μg/ml vs. >150 μg/ml) [17]. While cross resistance between RMP and other rifamycins, such as rifabutin and rifapentine, has been recorded [18], the compound structure of the drugs is different. This leads to subtle interaction differences between the binding site and the drugs, and could explain differential mutations causing resistance [19]. Further investigation using similar protein modelling approaches could shed light onto the mechanism of resistance to these drugs and highlight the key residues required for resistance.

INH is a compound that inhibits mycolic acid biosynthesis by binding to an enoyl-acyl carrier protein reductase encoded by the *inhA* gene. It is a pro-drug, which is activated by a catalase-peroxidase enzyme encoded by *katG*. Mutations in *katG* are more tolerated than those in *inhA*, and more frequent in drug sensitive isolates. The *katG* 315 mutations S315N/T account for the majority (60-80 %) of the INH resistance in clinical isolates [20]. Mutations affecting *inhA* usually appear in the promoter region of its operon (denoted *inhA*-promoter), leading to increased transcription. Whilst mutations in *katG* lead to moderate to high levels of resistance (always >1 mg/L), those affecting *inhA* confer a lower level of resistance [20] (<1 mg/L), and therefore if detected could allow INH to play a further role in treatment [21]. Computational initiatives involving protein structure modelling have been applied to understand better the molecular mechanisms of drug resistance, especially where multiple mutations are present. It has been established that the binding affinity of RMP-*rpoB* is most altered by common S450L and H445Y mutants, leading to less effective binding and resistance [22]. Similarly, the S94A mutant leads to decreased affinity of the drug on INH-*inhA* binding, and increased resistance [23].

SM is an aminocyclitol glycoside that binds to 16S rRNA and inhibits protein synthesis. Mutations in the S12 ribosomal protein encoded by *rpsL* have been linked to resistance. These mutations change the tertiary structure of the 16S rRNA leading to decreased affinity to

SM and high-level resistance. The majority (54 %) of SM resistance in clinical isolates has been attributed to the K43R mutation in *rpsL* [24]. Whilst mutations in *rpsL* confer a high level of resistance [25], those in *rrs* (encoding 16S rRNA) are thought to contribute to moderate levels of resistance [24, 26], and those in *gidB* confer low levels of resistance [27, 28]. EMB is a first line drug targeting arabinan synthesis, which affects the mycobacterial cell wall. It targets members of the *embCAB* operon, which code for arabinofuranosyl transferases involved in synthesising components of the cell wall. Mutations in *embB*, especially at codons 306, 406 and 497, are frequently observed and give rise to a low level of resistance [29]. The observed range of low to moderate resistance is mutation-specific [30] and thought to differ from other drugs in that it is more a step-wise process, with each mutation increasing the level of resistance [29]. Mutations in *embCAB*, *ubiA*, and *aftA* are thought to accumulate and can cause high levels of resistance observed in some clinical isolates [29].

To improve knowledge of genetic determinants of drug resistance, the use of whole genome association methods has been suggested [31]. Here we undertook whole genome analysis of 144 clinical isolates in the collection of the Special Programme for Research and Training in Tropical Diseases (TDR) [32], for which live material is available to the research community (http://bccm.belspo.be). The isolates were sourced from the TDR strain bank and were selected to encompass diverse geographical settings representing the four major *M. tuberculosis* lineages [33], as well as include susceptible and resistance strains within lineage. Drug susceptibility testing was performed using RMP, INH, EMB, SM, kanamycin (KAN), capreomycin (CAP), ethionamide (ETH), ofloxacin (OFL), and para-aminosalisylic acid (PAS). No testing was performed for pyrazinamide (PZA). The completeness of phenotypic MICs was highest in first-line drugs. A genome-wide association approach was used on 127 isolates to detect genetic variants associated with drug resistance. Typically, failing to account for population structure, in particular the phylogenetic- or lineage-related clustering, potentially involving outbreaks, may lead to false positive associations. Adjusting for principal components and removing lineage-informative mutations in regression analyses have been used to control for these confounding effects. The use of mixed regression models, which include a SNP-based estimate of between sample kinship as a random effect, is considered a more robust approach for isolates that are highly related or with familial relationships [34]. Application of these approaches identified established resistance loci [35]. Many of the loci were supported by evidence of evolutionary convergence, that is, the repeated and independent emergence of mutations in phenotypically resistant strains, identified as homoplastic SNPs in a phylogenetic tree [36].

Mutations in coding regions can have many different effects on a protein structure and function [37–40]. Structural bioinformatics approaches for modelling and mutation analysis were applied to the polymorphisms identified in the *rpoB* and *katG* genes. The effect of mutations on protein stability and interactions was assessed quantitatively with well-established computational methods, shedding light on molecular mechanisms giving rise to observed drug resistance. Whilst second-line drug resistance was tested for only 40 isolates - not sufficient to perform a genome-wide analysis - a number of novel mutations in candidate genes were identified.

Methods

Isolates and phenotypic methods

Susceptibility testing was performed in the Antwerp laboratory where the samples were stored as part of the Special Programme for Research and Training in Tropical Diseases (TDR) strain bank [32]. Isolated Mtb strains were previously collected from various geographical sites to create a diverse collection of well characterised drug resistant strains to provide a resource for the TB research community [32]. Single colonies were selected and kept on Löwenstein-Jensen (LJ) culture for drug susceptibility testing. Resistance patterns for the first line drugs were determined using the proportion method, with the critical concentrations 0.2 μg/ml INH, 40 μg/ml RMP, 4 μg/ml SM, and 2 μg/ml EMB. MIC were also investigated on LJ for RMP (10, 20, 30, 40, 80, and 120 μg/ml), INH (0.05, 0.2, 0.8, 1.6, and 3.2 μg/ml), SM (1, 2, 4, 8, and 16 μg/ml), and EMB (1, 2, 4, and 8 μg/ml). The critical thresholds of MIC for calling resistance were 0.2, 2, 4, and 40 μg/ml for INH, EMB, SM, and RMP, respectively [32]. The MIC values were discretised into three groups (susceptible, intermediate, and fully resistant) using natural cut-offs in their empirical distributions.

For the second line drugs PAS was tested on LJ at 0.5 μg/ml. The other drugs were tested on Middlebrook 7H11 agar at the following concentrations: OFL 2 μg/ml, KAN 6 μg/ml, CAP 10 μg/ml, and ETH 10 μg/ml. The proportion method was used for all second line drugs with a critical proportion of 1 %. Lyophilised isolates were sent to the London laboratory where they were grown on LJ prior to DNA extraction using the Bilthoven RFLP methodology [41].

Sequence data and variant calling

All DNA samples underwent Illumina sequencing on the HiSeq 2000 platform at the KAUST genomic facility, generating paired-end reads of 150 bp (Additional file 1: Table S1, pathogenseq.lshtm.ac.uk/tdr, Additional file 1: Table S2). All raw sequence data can be downloaded from the ENA short read archive (accession number PRJEB11653). For the raw sequence data, *trimmomatic* (v0.33) software [42] (parameters: LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:36) was used to remove or truncate reads of low quality. High quality reads were then mapped to the H37Rv reference genome (Genbank accession: AL123456.3) using the *BWA-mem* (v0.7.12) algorithm [43] (parameters: -c 100 -M -T 50). From the resulting alignments, *SAMtools* (v1.3) [44] and *GATK* (v3.5) [45] software (default parameter settings) were used to call SNPs and small indels, and the interaction of variants between the methods retained. Mappability values were calculated along the reference genome using *GEM-Mappability* software with a *k-mer* length of 50 bp and a 0.04 % substitution threshold [46]. Non-unique SNP sites (mappability values greater than one) were removed. Sample genotypes were called using the majority allele (minimum frequency 75 %) in positions supported by at least 20-fold total genome coverage, otherwise they were classified as missing. Isolates or SNPs with in excess of 10 % missing genotype calls were excluded. The final dataset included 144 isolates and 17,952 genome-wide SNPs.

Population structure and association analysis

The best-scoring maximum likelihood phylogenetic tree rooted on *Mycobacterium canetti* was constructed by *RAxML* (v8.2) software [47] (parameters: -T 10 -f a -x 12345 -m GTRGAMMA -p 12345 -N 100) using the 17,952 high quality SNP sites. *M. canetti* is a predecessor of *M. tuberculosis* and therefore provides a convenient root to map for both ancient and modern strains. Spoligotypes were inferred *in silico* using *SpolPred* [48] and matched perfectly with available experimental results. Strain-types were determined using lineage-specific SNPs [33]. Further population structure assessment was performed using principal components analysis [49], leading to covariates for adjustment in association analyses. Logistic regression models were employed to estimate the strength of association between the binary drug resistance outcome (resistance vs. susceptible) and the aggregate number of mutations by coding region, RNA loci, and intergenic regions, as well as operons. Similarly, proportional odds models were applied to a trichotomous phenotype based on MIC values (susceptible, intermediate and full resistance). As expected a number of genes would be reported as significant due to a large amount of cross-resistance between drugs, and we adjusted for the presence of other resistance in the regression models. The main association analysis using mixed models with a SNP inferred kinship matrix as a random effect was implemented in *EMMA* (v.1.1.2) [34]. The operons or functional units containing clusters of genes under the control of the same promoter were determined from *TBDB* [50]. Gene function was extracted from *Tuberculist* [51]. Permutation tests based on resampling MIC values were performed to establish a statistical significance cut-off for each drug to account for false positives arising from multiple locus tests. The established cut-offs were RMP 1.58×10^{-5}, INH $1.67 \times 10^{-5}_{,}$ SM 2.73×10^{-5}, and EMB 1.77×10^{-5}. All statistical analyses were performed using *R* (v3.2) software. To identify SNPs enriched by convergent evolution, the *phyC* approach [36] was employed using an available implementation [52].

Protein mutation modelling

An *apo* crystal structure for *katG* (1SJ2 [53]) was available and downloaded from the Protein Data Bank (*PDBe* [54]). A protein homology model for *rpoB* was obtained from the Chopin database (http://mordred.bioc.cam.ac.uk/chopin). Reliable models could not be found or generated for *embB*, *rpsL* or other loci identified in our work. Structures of the drug compounds INH and RMP where obtained from the chemical components section of PDBe and used in *Autodock vina* [55] to perform *in silico* drug docking. The *mCSM (*http://structure.bioc.cam.ac.uk/mcsm*)* and *DUET (*http://structure.bioc.cam.ac.uk/duet*)* web servers were used to assess changes in protein stability and mCSM-PPI (http://bleoberis.bioc.cam.ac.uk/mcsm/protein_protein) to quantify effects on protein-protein interactions [56, 57].

Results

Genetic polymorphisms

The 144 isolates represented a broad global distribution, sourced from 24 countries in four continents (Additional file 2: Figure S1, Additional file 1: Table S1). All the African isolates were lineage 4 strains, and only Asia contributed lineage 1 strains. Across the isolates, 19,248 SNP sites were identified, including 17,092 (89 %) in coding regions of the genome (11,003 [(57 %] non-synonymous mutations). The SNP allele frequency spectrum revealed, as expected, the majority of variants were rare (12,244 [63.5 %] SNPs present in only one isolate; Additional file 3: Figure S2). Both a phylogenetic tree and a principal component analysis based on the ~19 k SNPs showed congruent clustering by lineage (Additional file 4: Figure S3). The tree revealed a cluster of nine Rwandan strains, which were separated by low numbers of SNP differences (range 1-17 SNPs), implying potential transmission. It also revealed one sample reported as susceptible to EMB was likely to be resistant due to its location on the tree within a cluster of isolates with resistance.

Drug resistance

The drug susceptibility test MIC values for the four first line drugs were available for 144 isolates, and 17 strains were removed due to poor sequence coverage and quality. For the remaining 127 isolates, similar numbers of sensitive and resistant strains were present (Fig. 1). For

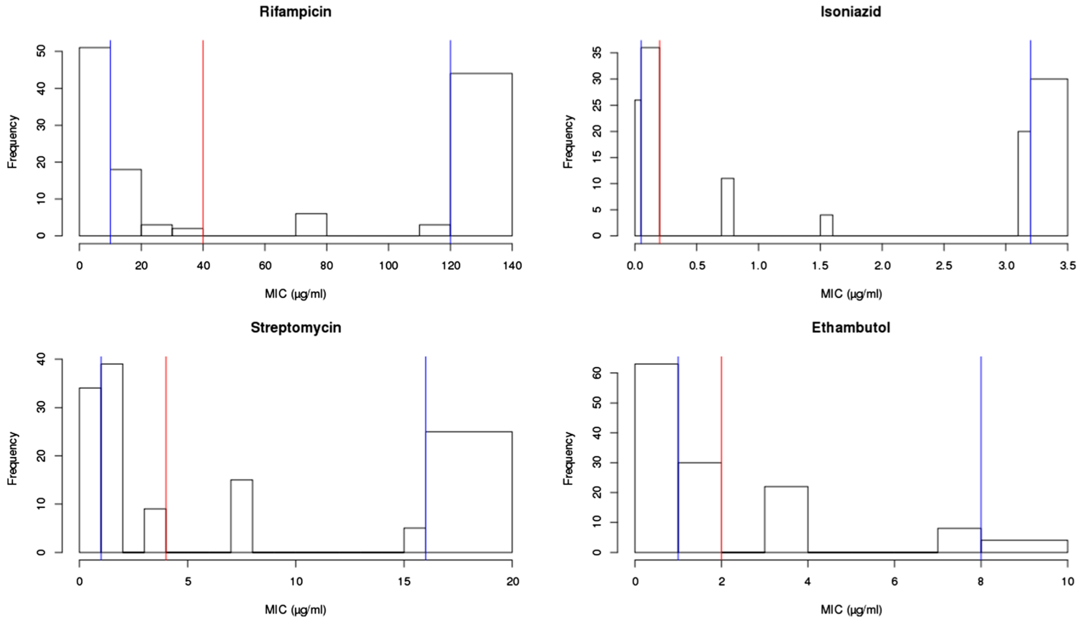


Fig. 1 The distribution of MIC values for rifampicin, isoniazid, streptomycin, and ethambutol. The *red vertical line* is the standard susceptible-resistance threshold (rifampicin 40 µg/ml, isoniazid 0.2 µg/ml, streptomycin 4 µg/ml, ethambutol 2 µg/ml). The two *blue vertical lines* define the three levels (susceptible, intermediate and full resistance): rifampicin (10, 120 µg/ml), isoniazid (0.05, 3.2 µg/ml), streptomycin (1, 16 µg/ml) and ethambutol (1, 8 µg/ml). *MIC* minimum inhibitory concentrations

the trichotomised MIC values, the intermediate resistance group comprised less than 20 % of isolates across drugs (see Fig. 1 for breakpoints). There was a high correlation between INH and other drug MIC values (Spearman's *rho* >0.31, p <0.006), and in total there were 14 distinct drug resistance combinations across the four first-line drugs, in keeping with the step-wise and combination nature of therapies. Twelve (9.4 %) isolates were pan-resistant, 38 (29.9 %) pan-susceptible, and 42 (33.1 %) multi-drug resistant (using dichotomised values, Additional file 1: Table S3). The *TB profiler* [11] was used to infer drug resistance profiles *in silico* from known drug resistance mutations. Assuming the drug susceptibility tests as the reference standard, the computationally inferred resistance profiles were highly accurate for RMP (sensitivity/specificity: 0.962/1.000) and INH (0.908/0.935), suggesting the sequencing result would be of clinical value for detecting MDR-TB. The performance for SM (sensitivity/specificity: 0.511/0.960) and EMB (0.971/0.839) was less accurate. High predictive values will be needed to guide the use of SM and EMB in patients with MDR and XDR-TB. It would appear that the repertories of mutations and loci for these drugs still need to be elucidated and that intermediate resistance with MIC values close to the resistance cut-offs could pose problems using binary outcome values when correlating genotype and phenotype. Mutations in the *gid* gene are not included in *TB Profiler* as they cause only intermediate levels of SM resistance. We observed twenty *gid* markers and their incorporation increased the SM sensitivity to 82 %. Further, it was predicted that 14 (11 %) isolates were likely to be PZA resistance. In particular, each of the 14 isolates had at least one known drug resistance conferring mutation in the *pncA* gene (Ala171Pro, Arg121Pro, Asp8Ala, Gln10Pro, His57Pro, His82Asp, Ile31Ser, Ser66Pro, Thr76Pro [n = 2], Trp68Ser, Tyr103His, and Val125Gly [n = 2]).

In an attempt to search for new mutations involved in drug resistance a genome wide association analysis was performed on both trichotomous MIC and binary resistance phenotypes. Both single SNP and locus-wide association testing were considered. Similar to a rare variant analysis, the number of (non-synonymous) mutations per sample, per gene and operon was calculated, and correlated with the phenotype. In addition to association analysis, the complementary *phyC* approach was applied. This approach aims to identify loci under convergent evolution in resistance branches of the tree. A summary of all statistically significant results is presented (Table 1), and we focus on each drug separately.

Table 1 First-line drug related SNPs identified in association and convergent evolution analysis

Drug	Gene	SNP mutations (% in resistant isolates)
Rifampicin	*rpoB*	T400A (3.8), D435V (9.4), H445D/Y (11.3),
		H445R (5.7), **S450W/L** (60.4), I491V/F (3.8)
Isoniazid	*katG*	**S315N** (69.2)
Isoniazid	*Rv1482c-fabG1 (inhA*-promoter)	**C-15 T** (24.6)
Streptomycin	*rpsL*	**K43R** (24.4)
Ethambutol	*embB*	C12T (5.9), M306I (14.7*), **M306V** (17.7*), D354A (11.8), G406S/C (11.8), G406D/A (11.8**), **Q497P/R** (17.7***), D1024N (8.8)
Ethambutol	*cadI*	**C-39 T** (8.8)

The genes were identified using aggregated mutation mixed models. The SNPs were identified using the *phyC* method and those also found using the GWAS mixed model approach are highlighted in **bold**

SNP single nucleotide polymorphism, *GWAS* genome-wide association study

*observed in "sensitive" strains at frequency 3.2 %; **4.3 %; ***1.1 %; all $P < 1 \times 10^{-5}$ from association analysis

Rifampicin

Genome-wide analysis using both binary trait or MIC values revealed, as expected, that the *rpoB* gene ($p < 1 \times 10^{-20}$) and its operon ($p < 1 \times 10^{-10}$) were associated with RMP resistance. One tri-allelic SNP in *rpoB* at position 761,155 (codon 450: S450L 30/127, S450W 2/127) was associated with the majority of RMP drug resistance (60 %). There were six significant SNPs under convergent evolution ($p < 0.05$) in *rpoB* (codons 450, 445 (*x*2), 435, 400, and 491), one in *rpoC* (N416S mutation, two isolates, a known compensatory mechanism) and one in *lldD2* (codon 2 synonymous, 16 isolates). Fifty isolates (93 % of RMP resistant strains) had at least one mutation in the *rpoB* gene in the RMP resistance determining region (codon range 400-491) (Fig. 2a). Three isolates had two mutations in this region. Two isolates had mutations in codons 400 and 450 and one strain had mutations in codons 450 and 491. All except four isolates with a mutation in *rpoB* had MIC values of at least 120 μg/ml and the remaining four had values of 80 μg/ml.

Isoniazid

The association analysis revealed the *Rv1907c-furA* operon ($p < 1 \times 10^{-13}$), which contains the *katG* gene ($p < 1 \times 10^{-9}$) as the most significant association (Fig. 2b). Other loci identified included the *fabG1-hemZ* operon (contains the *inhA* gene and promoter). Using MIC values, the *Rv1907c-furA* ($p < 2 \times 10^{-5}$) operon and *katG* and *Rv1979c* genes were found to be associated with INH resistance. A SNP-based GWAS revealed a single polymorphism association in *katG* (position 2,155,168, S315T/N, $P < 4.33 \times 10^{-18}$). This SNP was supported by *phyC* analysis, which also revealed another site under convergent evolution in *inhA* promoter. Overall, 47 (75 % of INH resistant) strains have a SNP in position 2,155,168 (S315T 41 isolates, S315N four strains), of which 43 have an MIC value of at least 3.2 μg/ml, while the remaining two had values of 0.8 and 1.6 μg/ml. Twenty-one isolates have a SNP in the *fabG1-hemZ* operon, with MIC values ranging from 0.8 to ≥3.2 μg/ml. Of the 16 isolates that only have one SNP in the *fabG1-hemZ* operon, half had MIC values in excess of 0.8 μg/ml. The three isolates with mutations at both the *fabG1* promoter and *inhA* had an MIC value in excess of 1.6 μg/ml. Three (of six) isolates with a mutation in the promoter and an MIC of at least 3.2 μg/ml also have the *katG* S315T mutation. One mutation in the *katG* promoter region was found in a drug sensitive sample.

Streptomycin

The association analysis identified the *rpsL-rpsG* operon and the *rpsL* gene as being associated with SM resistance (Fig. 2c). The *rpsL* locus was also found by analysing MIC values, and a SNP-based approach identified one mutation (position 781,687, K43R, 11 isolates, 26 % of resistant strains) within the gene. The *phyC* method identified two SNPs in the rRNA gene *rrs* (514 A- > C, four isolates; 517 C- > T, three strains). All isolates, except one, had an MIC of greater than 16 μg/ml. One sample with the 1,472,362 C- > T mutation had an MIC of 8 μg/ml.

Ethambutol

A binary phenotype analysis identified the *embA-embB* operon ($p < 1 \times 10^{-10}$) and the *embB* gene ($p < 1 \times 10^{-13}$) (Fig. 2d). This result was confirmed in an analysis of the MIC phenotype (operon $p < 1 \times 10^{-10}$, gene $p < 1 \times 10^{-8}$). A SNP-based association analysis revealed one in the *embB* gene (position 4,248,003) and one in the promoter of *cadI,* where the latter was also found using the *phyC* method (four isolates) (Fig. 2e) The *phyC* approach identified seven SNPs in *embB* (codons 306 [*x*2, 22 isolates], 354 [four resistant isolates], 406 [*x*2, 12 isolates], 497 [seven isolates], and 1024 [two isolates]). Three isolates had mutations in two of these positions and all others had only one mutation. There was a great range of MIC values in isolates containing these mutations with some codons having both sensitive and resistant strains. For

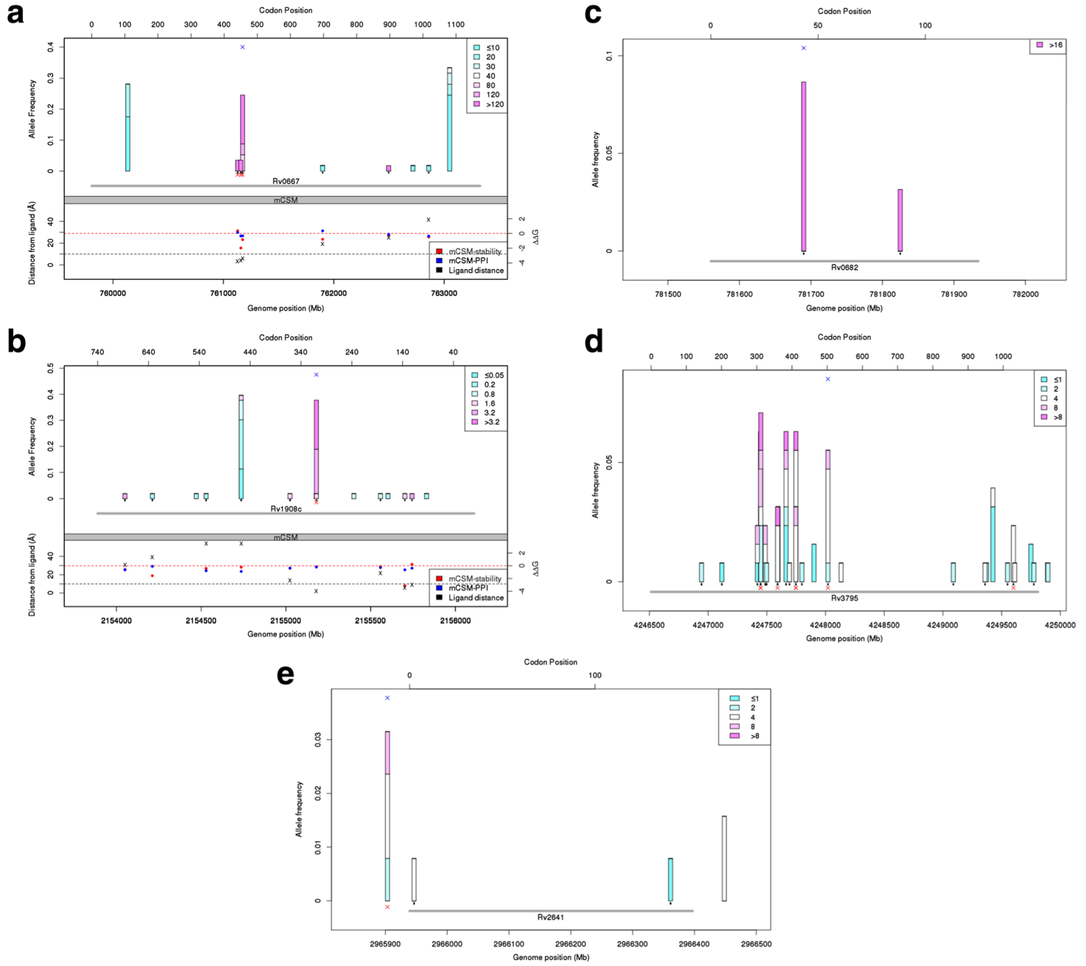


Fig. 2 SNPs in candidate genes in isolates with a single mutation in each locus. The *bars* represent the allele frequency of the SNPs, and are coloured according to the MIC value. *Black dots* under bars represent non-synonymous mutations. *Blue* and *red crosses* represent mutations that have been found to be significant in the association and the convergent evolution *phyC* analyses, respectively. Structural data are available only for *rpoB* and *katG* (*bottom panels*). The protein stability and protein-protein interaction changes induced by the SNP as calculated by *mCSM* software are represented by the *red* and *blue points*, respectively, and magnitude is represented on the *right y-axis*. The distance of each mutated codon from the docked drug *(left y-axis*) is denoted by the *black crosses*. **a** Rv0667 *rpoB* (rifampicin). **b** Rv1908c *katG* (isoniazid). **c** Rv0682 *rpsL* (streptomycin). **d** Rv3795 *embB* (ethambutol). **e** Rv2641 *cadI* (ethambutol). *SNPs* single nucleotide polymorphisms, *MIC* minimum inhibitory concentration

example, 6/22 isolates with mutations in codon 306 had MIC values of at most 2 μg/ml. Mutations in the *embA* promoter were also present, but not found to have a consistent effect on the MIC values when combined with mutations in *embB*. The additive effect of mutations in the candidate genes *embB*, *embA*, *embA* promoter, *embC*, *embR*, and *ubiA* correlated modestly with MIC values (*rho* = 0.24, Additional file 5: Figure S4). The aggregated mutation approach revealed that the *pncA* gene may be associated with EMB resistance, but this was most likely due to cross-resistance from the predicted PZA resistant cases (n = 14).

Use of MIC values

The correlation between association p-values using binary resistance (susceptible, resistant) and trichotomous MIC was modest (RMP 0.386, INH 0.311, EMB 0.309, and SM 0.360), but led to near identical strongest hits. However, there were some discrepancies in the findings for EMB and SM. The majority of isolates (11/15) that were EMB phenotypically susceptible, but with known drug resistance mutations, had an MIC value of 2 μg/ml. This value is on the upper bound of the sensitive range, but low-level resistance may be predicted as they had known EMB drug resistance mutations. The majority of SM false negative (15/22) isolates had an MIC value of 8 μg/ml, which is on the lower limit of the resistance cut-off. Mutations in *gid* are known to cause low levels of resistance, and the majority (19/22) of false negative strains contained mutations in that gene. The additive effect of mutations in both EMB and SM candidate genes correlated with increasing MIC value (EMB: *rho* = 0.24, slope = 0.29, p = 0.003; SM: *rho* = 0.48, slope = 3.59, p = 1.65×10^{-8}; Additional file 6: Figure S5), and may provide some evidence of accumulating low resistance mutations.

An exciting prospect is the use of MIC values to infer the additive and interaction effects of each mutation. Unfortunately, the relatively small sample size did not allow a rigorous statistical approach to look for interactions. However, the frequencies of combinations of mutations for RMP, INH, EMB, and SM, and their MIC values are presented (Additional file 1: Table S4). Using these data, statistical models were fitted with all mutations included, to allow an assessment of the MIC variation explained and their independent effects in the presence of others (Additional file 6: Figure S5). For RMP and INH, a high proportion of MIC variation is explained by single mutations (RMP: *rpoB* 450, 48.4 %, INH: *katG* 315, 73.8 %). However, for EMB and SM, single mutations explained at most ~30 % (SM: *rpsL* codon 43 – 32.4 %, EMB: *embB* codon 306 – 30.0 %), with the largest proportion due to unknown factors (SM: 44.0 %, EMB: 37.4 %). This analysis further supports that other variants need to be identified for EMB and SM drugs.

We compared the association results from the mixed models using all available data to regression-based approaches that adjusted for the principal components (explained ~60 % of variation) and removed 414 lineage- and clade-specific markers and eight highly related Rwandan strains (Additional file 4: Figure S3). There was a moderate level of correlation between the approaches for all outcomes (minimum *rho* - RMP: 0.66, INH: 0.54, SM: 0.20, EMB: 0.34). This correlation translated into identical top hits for association (Table 1), except for the *cadI* gene, which was identified only by the mixed model approach at the stringent significance cut-off. CadI is a protein that can be induced by cadmium, and is thought to possess similar functions to the metallothioneins and protects the bacterium against metal toxicity (http://tuberculist.epfl.ch).

Second-line drugs

Forty-four (35.8 %) isolates were tested for second line drug resistance, and the polymorphism in known candidates was considered (Table 2). Of the six isolates that were resistant to PAS, mutations at candidate genes (*folC, ribD, thyA*, and *thyX*) were observed in all isolates (*folC* E40G, I43G, D135G; *thyA* Y94C, Q97R, V135F; and *thyX* promoter G-16A (n = 2), T-43G). Seven isolates had ETH resistance, of which all had mutations in drug resistance candidate genes (*ethA* R469P, n = 1; *ethR-fabG1* promoter region C-15 T, n = 6; and *inhA* gene S94, n = 1). Three isolates had resistance to OFL, with known mutations in the *gyrA* gene (D94G, n = 2; N499D, n = 1). Two isolates had resistance to CAP, with unreported mutations in candidate genes (*rrs* A1205G, n = 1; *tlyA* gene G196E, n = 1). No indels were identified in these genes.

Effects on protein structure and function

The availability of structural information for *katG* and *rpoB* genes allowed us to assess the potential functional effects of the mutations identified and their ability to predict drug resistance. The respective INH and RMP drugs were computationally docked into the models, delimiting the residues of the drug binding site. The *mCSM* and *DUET* servers were used to quantify the influence of mutations on protein stability and protein-protein interactions (measured by the change in Gibbs free energy ΔΔG between the wild-type and mutant structures). These factors, individually or combined could lead to drug resistance. The predictions obtained are summarized in Additional file 1: Table S5.

Across the eleven RMP resistance codons analysed in *rpoB* and ten INH resistance codons of *katG*, no strong correlation of the changes in protein stability with the proportion of drug resistant isolates with each mutation was observed (*rho* < 0.05, p >0.05). There was weak evidence that drug resistant isolates had mutations that were more destabilizing (p <0.10). The mutations in *katG* were not located near the homodimer interface, while further structural information is necessary to characterise the *rpoB* interactions. However, across both drugs there was a strong association between (a shorter) distance of the mutation to the ligand in the protein structure and resistance

Table 2 Second-line drug related mutations in candidate genes

Drug	No. resistant	Locus (codon [no. isolates])
Para-aminosalisylic acid	6	*folC* (E40G[1], I43G[1], D135G[1]);
		*thyA (***Y94C**[1], Q97R[1], **V135F**[1]);
		thyX promoter (**G16A** [2], **T43G** [1]).
Ethionamide	7	*ethA* (**R469P**[1]);
		ethR-fabG1 promoter (C15T[6]);
		inhA (S94[1])
Ofloxacin	3	*gyrA (*D94G [2]; N499D [1]).
Capreomycin	2	*rrs* (**A1205G**[1]); *tlyA* (**G196E** [1])

Previously unreported in **bold**

(greater MIC values) (*rpoB rho* = -0.79, p = 8.1 × 10^{-6}; *katG rho* = -0.72, p = 0.0012) (Fig. 3). For RMP, isolates with MIC values of at least 80 μg/ml had mutations located close to the drug binding site (median distance of 5.77 Å, all values less than 10 Å) as depicted in Fig. 4, compared to isolates with MIC values of ≤10 μg/ml (median distance of 37.08 Å). For INH, isolates with MIC resistance values of at least 3.2 μg/ml had mutations directly interacting with the drug (median 2.15 Å) (Fig. 4), whilst isolates with intermediate resistance (1.6 μg/ml) mutations located further away (median 9.93 Å), and mutations in susceptible strains (MIC values less than 0.8 μg/ml) were even more remote (median 53.97 Å). Additional file 7: Figure S6 shows the molecular interactions established by mutated residues in *katG* and *rpoB*, with most of the effects of mutations influencing interactions established directly with the drug molecule, by destabilizing the surrounding region via loss of interactions or the introduction of steric clashes. Whether we can predict the resistance of a mutation using its distance to a ligand site will have to be verified using other protein structure models, when they become available.

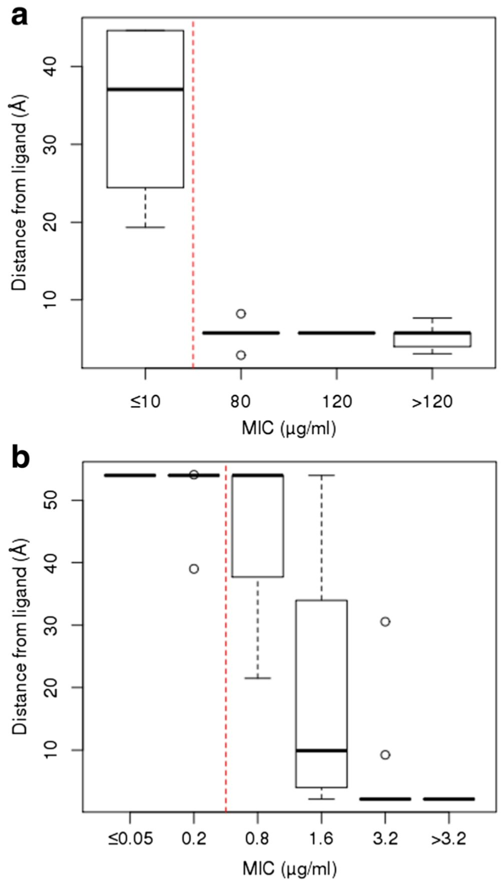


Fig. 3 Boxplot showing the distributions of the distance of the mutated codon to the drug for all the SNPs in each MIC level in (**a**) RMP-*rpoB* and (**b**) INH-*katG*. *Vertical red line* is the resistance cut-off. *SNP* single nucleotide polymorphism, *MIC* minimum inhibitory concentration, *RMP* rifampicin, *INH* isoniazid

Discussion

Early characterisation of drug resistance mutations would assist TB patient management and avoid treating individuals with inefficacious toxic regimens [11]. Current testing for resistance to most anti-tuberculosis drugs, as applied to isolates in TDR, involves isolation and culture of the bacteria followed by exposure to the drug, a process that takes weeks or months [11]. However, the direct sequencing of *M. tuberculosis* from sputum from suspected drug resistant patients [58] and the development of rapid strain profiling tools, suggests that culture-free approaches have a role in the management of TB [11]. For some drugs, such as RMP and INH, resistant mutations are well characterised, but for others such as SM, EMB and second-line treatments, existing databases lack specificity and sensitivity [11]. We performed a genome-wide association approach on SM and first-line treatments and assessed its ability to confirm existing, and identify new, variants that cause drug resistance. Whilst genome-wide association methods have become established for disease susceptibility studies in humans, their application in pathogens is still in its infancy [31]. Population structure can confound analyses and lead to false positive results. For TB, widespread drug resistance may be over represented in particular lineages or clades, causing lineage specific SNPs that confound analyses. This confounding was handled by a mixed model, but alternative approaches were considered, in particular, removal of all lineage- and clade-specific markers or inclusion of principal components as surrogates for lineages within the regression model. These approaches led to near identical top association hits, in part reflecting the strong signal of the resistance-related mutations across clades, the dominant clustering of discrete lineages in the phylogeny, and the modest number of highly related or outbreak-based isolates (e.g. Rwandan strains). Our work suggested that the use of kinship matrices within mixed models may avoid the removal of lineage-informative SNPs and highly related strains, especially those involved in an outbreak or transmission study. This observation is supported in human GWAS studies with familial relationships, where mixed models have been found to be more robust to false positive associations than principal components adjustment [59].

A limitation of the study was the representation of geographic origins and lineages, as we were restricted by availability of strains collected for this extremely well

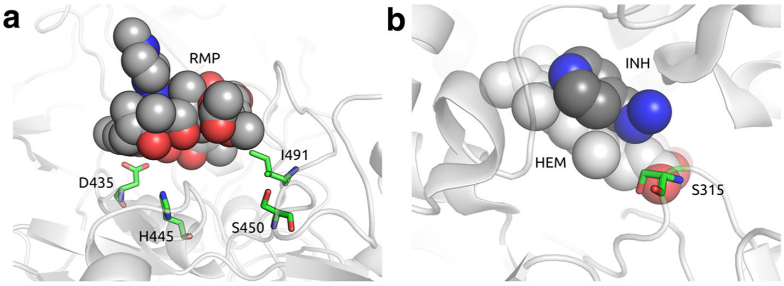


Fig. 4 Mutations in binding site regions. **a** depicts the spatial distribution of mutated residues in the *rpoB*-RMP complex while (**b**) shows the residue Ser315 in *katG*-INH complex (residues depicted with carbons in green). The distance between the residues and the ligands (depicted with carbons in dark grey) vary from 2.1 to 5.7 Å. *RMP* rifampicin, *INH* isoniazid

characterized collection. A second limitation was the small sample size, especially for analyses of second-line drugs, where a genome-wide approach could not be implemented. However, where sample sizes were sufficient our genome-wide analysis reported genes known to be involved in first-line RMP, INH, SM, and EMB drug resistance. The use of MIC values has been advocated as a more sensitive measure, but the potential lack of a symmetric distribution of values (as shown in our data) could lead to invalidation of assumptions for linear models. We took the pragmatic approach of discretising the values into three natural groups (resistant, sensitive, and intermediate) allowing an alternative modelling strategy (proportional odds model) to be employed. The correlation between association analysis p-values using both binary and trichotomised MIC values was modest (range: 0.31-0.39). Some isolates with intermediate SM resistance had no known drug resistance mutations in *rpsL* and *rrs*, and even after inclusion of *gid* mutations, additional causal mutations or genes to explain phenotypic variation remained unidentified. Larger sample sizes would facilitate the use of raw MIC values and therefore advance the detection of variants that confer intermediate resistance. Many of the results were also confirmed using convergent evolution methods, which require smaller sample sizes than genome-wide approaches, and should prove to be a powerful and robust method to detect drug resistance mutations in *M. tuberculosis*, and possibly other pathogens. There are a number of isolates that have very high levels of resistance to both EMB and SM but do not present any mutations in known candidate genes. It is evident that there are rare SNPs occurring in unknown genes that confer EMB resistance. Similarly, there are many isolates with more than one mutation in candidate genes and high levels of susceptibility. Not all mutations in these genes will have an effect on resistance levels, and interactions between the drug and its target should be considered.

The use of protein structures determined by X-ray crystallography or as homology can provide extra validation and an insight into the mechanism of drug resistance conferred by mutations. It has been shown that mutations in the RMP binding site can cause resistance due to disturbance of the active site both in Mtb and in other bacteria [22]. An exciting finding was the strong correlation between the MIC values and the distance in the three-dimensional structure of the mutated residue to the drug docking ligand. This observation seems novel to Mtb. If it holds for other genes as their protein structures become available, then potential drug resistance mutations could be predicted *in silico* in a genome-wide screen. The binding sites of the rifamycins have been shown to be in similar locations and these observations would be expected to be similar for closely related drugs [60]. It could also provide a future high throughput way of integrating genomic and protein structure data to make predictions about drug resistance mutations. In particular, rare SNPs with low allele frequencies may not be detected in association analyses; however, prediction of the distance of the mutated codon to a ligand or its effect on overall stability or protein-protein interactions can provide a complementary approach to identify new drug resistance conferring mutations. Indeed, variants such as the *rpoB* V170F mutant are present in only one isolate in our dataset but it was flagged up as an interesting SNP due to its proximity to the docked RMP ligand in the homology model. This *rpoB* SNP has been attributed to drug resistance by earlier studies[12].

Conclusions

Overall, our work has demonstrated the potential of the genome-wide association and selection approaches to identify mutations and genes associated with resistance. We have also shown that if protein structures are available, then the effects of mutations in genes on resistance may be predicted *in silico*. This could facilitate the prediction of the effects of mutations on novel drugs and potential resistance. Ultimately, such insights will assist with patient treatment and management, and disease control.

Availability of data and materials

All raw sequence data can be downloaded from the ENA short read archive (accession number PRJEB11653).

Additional files

Additional file 1: Table S1. The isolates according to geographic location and phenotypic drug resistance. CAR Central African Republic; DRC Democratic Republic of Congo, L1-L4 lineages 1 to 4, (first line drugs) RMP = rifampicin, INH = isoniazid, SM = streptomycin, EMB = ethambutol; (second line drugs) OFL = ofloxacin, KAN = kanamycin, CAP = capreomycin, Et = ethionamide, P = Para-aminosalisylic acid. **Table S2**. The isolate ENA accession numbers and MIC values. RMP rifampicin, INH isoniazid, SM streptomycin, EMB ethambutol. **Table S3**. Drug susceptibility profiles for rifampicin, isoniazid, streptomycin and ethambutol. R = resistance, S = sensitive; 13 different profiles were identified across 127 independent isolates; Multi-drug resistant in italics. **Table S4**. Combinations of mutations and their frequency (N) in drug resistance candidate genes. a) Rifampicin. b) Isoniazid. c) Streptomycin. d) Ethambutol. * single mutation, ** double mutations, *** triple mutations; SNP mutations in a single sample have been aggregated into a "rare" column. **Table S5**. Predicted effects of mutations. (DOCX 55 kb)

Additional file 2: Figure S1. The global distribution of geographic origin and lineage of the isolates. Lineages one to four are represented by blue, green, purple, and red, respectively. (PNG 265 kb)

Additional file 3: Figure S2. SNP allele frequency spectrum. A large number of rare variants are observed. Peaks with higher allele frequency reflect the presence of lineage and sub-lineage specific SNPs. (PNG 33 kb)

Additional file 4: Figure S3. Population structure analysis of the 144 isolates show clustering by lineage (Lineages one to four are represented by blue, green, purple, and red points, respectively). (a) A phylogenetic tree rooted with *M. canetti*. (b) First two principal components represent 33 % and 30.5 % of the variation explained between isolates, respectively. (ZIP 105 kb)

Additional file 5: Figure S4. The relationship between the total number of non-synonymous SNPs in candidate loci and the MIC values. The size of the circle represents the number of isolates. a) Ethambutol (*embB, embA, embA promoter, embC, embR* and *ubiA*). b) Streptomycin (*rpsL, rrs*). The size of the circles is proportional to the frequency. The MIC values tend to increase with the number of non-synonymous mutations (ethambutol: *rho* = 0.24, slope = 0.29, p = 0.003; streptomycin: *rho* = 0.48, slope = 3.59, p = 1.65×10^{-8}). The horizontal blue lines refer to the resistance cut-offs. (ZIP 92 kb)

Additional file 6: Figure S5. Percentage of the variation in MIC values explained by each mutated codon in candidate genes. Bars in red represent significant independent associations with increased MIC (p < 0.05). a) Rifampicin. b) Isoniazid. c) Streptomycin. d) Ethambutol. (ZIP 231 kb)

Additional file 7: Figure S6. Molecular interactions established by wild-type residues in *katG* and *rpoB* residues. (A) The interactions established by Ser315 in *katG*. Given the proximity of the residue to the ligands INH and HEM, mutations to Asn and Thr, with slightly larger side chains, would potentially cause steric clashes. (B) The interactions of Asp435 in *rpoB*. It directly interacts with RMP via polar interactions that would be disrupted by mutations to Val. (C) Thr400 in *rpoB* is at the end of an alpha helix establishing intra molecular interactions. Giving its distance to RMP, it would be expected that its mutation to Ala would be a lower impact, which would arise from alosteric changes. (D) Ser450 establishes strong intra molecular interactions in the RMP binding site. Mutations to larger residues (Trp and Leu) could disrupt the packing of the region and therefore binding. (E). Ile491 performs hydrophobic interactions with RMP and its neighbouring residues. Mutations to Phe or Val would compromise packing, either inducing steric clashes or compromising packing. (F). His445 performs strong intra molecular interactions, including a donor-pi (*blue dashes*) and hydrogen bond (*red dashes*). Mutations to residues Asp, Tyr or Arg would imply in the loss of the pi interaction as well as potential introduction of steric clashes. (PNG 749 kb)

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

RM, NC, GAH, MBN, KM, AR, SC, AP, and LR contributed isolates, sequencing resources and/or reagents. JP, FC, DBA, DEVP, NF, MLH, and TGC contributed towards data analysis. JP, RM, DBA, DEVP, SC, and TGC contributed towards writing the first version of the manuscript, with all other authors contributing to the final version. All authors read and approved the final manuscript.

Acknowledgements

JP is supported by a BBSRC PhD studentship. The project was supported by the KAUST faculty baseline research fund (KAUST-BRF) to AP. The authors wish to thank members of KAUST Bioscience Core laboratory who sequenced the isolate DNA. DBA is supported by an NHMRC CJ Martin Fellowship (APP1072476). DEVP is supported by René Rachou Research Center (CPqRR/FIOCRUZ Minas). DBA and DEVP are funded by a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG). NF is funded by a Medical Research Council Methodology Research Fellowship (MR//K020420). TGC is funded by the Medical Research Council UK (Grant no. MR/K000551/1, MR/M01360X/1, MR/N010469/1). There are no conflicts of interests.

Author details

[1]Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK. [2]University of Cape Town Lung Institute, Lung Infection & Immunity Unit, Old Main Building, Groote Schuur Hospital, Observatory, Cape Town 7925, South Africa. [3]Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK. [4]Centro de Pesquisas René Rachou, Fundação Oswaldo Cruz, Avenida Augusto de Lima 1715, Belo Horizonte 30190-002, Brazil. [5]Mycobacteriology Unit, Institute of Tropical Medicine, Antwerp, Belgium. [6]Pathogen Genomics Laboratory, BESE Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. [7]Sydney Emerging Infections and Biosecurity Institute and School of Public Health, Sydney Medical School, University of Sydney, Sydney, NSW 2006, Australia. [8]Special Programme for Research and Training in Tropical Diseases (TDR), World Health Organisation, Geneva, Switzerland. [9]Department of Biomedical Sciences, Antwerp University, Antwerp, Belgium. [10]Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK. [11]Department of Pathogen Molecular Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London, UK.

Received: 7 December 2015 Accepted: 2 February 2016
Published online: 23 March 2016

References

1. World Health Organization. Global tuberculosis report 2014. Geneva; 2014.
2. Gagneux S. Host-pathogen coevolution in human tuberculosis. Philos Trans R Soc Lond B Biol Sci. 2012;367:850–9.
3. Galagan JE. Genomic insights into tuberculosis. Nat Rev Genet. 2014;15:307–20.
4. Clark TG, Mallard K, Coll F, Preston M, Assefa S, Harris D, et al. Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing. PLoS One. 2013;8:e83012.
5. Kent PT, Kubica GP. A guide for the level III laboratory. Atlanta: CDC; 1985.
6. Canetti G, Fox W, Khomenko A, Mahler HT, Menon NK, Mitchison DA, et al. Advances in techniques of testing mycobacterial drug sensitivity, and the use of sensitivity tests in tuberculosis control programmes. Bull World Health Organ. 1969;41:21–43.
7. Nebenzahl-Guimaraes H, Jacobson KR, Farhat MR, Murray MB. Systematic review of allelic exchange experiments aimed at identifying mutations that

confer drug resistance in Mycobacterium tuberculosis. J Antimicrob Chemother. 2014;69:331–42.

8. Reynolds MG. Compensatory evolution in rifampin-resistant Escherichia coli. Genetics. 2000;156:1471–81.
9. Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, et al. Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. Nat Genet. 2012;44:106–10.
10. de Vos M, Müller B, Borrell S, Black PA, van Helden PD, Warren RM, et al. Putative compensatory mutations in the rpoC gene of rifampin-resistant Mycobacterium tuberculosis are associated with ongoing transmission. Antimicrob Agents Chemother. 2013;57:827–32.
11. Coll F, McNerney R, Preston M, Guerra-Assunção JA, Warry A, Hill-Cawthorn G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med. 2015;7:51.
12. Witney AA, Gould KA, Arnold A, Coleman D, Delgado R, Dhillon J, et al. Clinical application of whole-genome sequencing to inform treatment for multidrug-resistant tuberculosis cases. J Clin Microbiol. 2015;53:1473–83.
13. Koch A, Mizrahi V, Warner DF. The impact of drug resistance on Mycobacterium tuberculosis physiology: what can we learn from rifampicin? Emerg Microbes Infect. 2014;3:e17.
14. Telenti A, Imboden P, Marchesi F, Lowrie D, Cole S, Colston MJ, et al. Detection of rifampicin-resistance mutations in Mycobacterium tuberculosis. Lancet. 1993;341:647–50.
15. Campbell EA, Korzheva N, Mustaev A, Murakami K, Nair S, Goldfarb A, et al. Structural mechanism for rifampicin inhibition of bacterial rna polymerase. Cell. 2001;104:901–12.
16. Brandis G, Hughes D. Genetic characterization of compensatory evolution in strains carrying rpoB Ser531Leu, the rifampicin resistance mutation most frequently found in clinical isolates. J Antimicrob Chemother. 2013;68:2493–7.
17. Jamieson FB, Guthrie JL, Neemuchwala A, Lastovetska O, Melano RG, Mehaffy C. Profiling of rpoB mutations and MICs for rifampin and rifabutin in Mycobacterium tuberculosis. J Clin Microbiol. 2014;52:2157–62.
18. Heep M, Beck D, Bayerdörffer E, Lehn N. Rifampin and rifabutin resistance mechanism in Helicobacter pylori. Antimicrob Agents Chemother. 1999;43:1497–9.
19. Sirgel FA, Warren RM, Böttger EC, Klopper M, Victor TC, van Helden PD. The rationale for using rifabutin in the treatment of MDR and XDR tuberculosis outbreaks. PLoS One. 2013;8:e59414.
20. Jacobson KR, Theron D, Victor TC, Streicher EM, Warren RM, Murray MB. Treatment outcomes of isoniazid-resistant tuberculosis patients, Western Cape Province. South Africa Clin Infect Dis. 2011;53:369–72.
21. Schönfeld N, Bergmann T, Vesenbeckh S, Mauch H, Bettermann G, Bauer TT, et al. Minimal inhibitory concentrations of first-line drugs of multidrug-resistant tuberculosis isolates. Lung India. 2012;29:309–12.
22. Kumar S, Jena L. Understanding rifampicin resistance in tuberculosis through a computational approach. Genomics Inform. 2014;12:276–82.
23. Wahab HA, Choong YS, Ibrahim P, Sadikun A, Scior T. Elucidating isoniazid resistance using molecular modeling. J Chem Inf Model. 2009;49:97–107.
24. Sreevatsan S, Pan X, Stockbauer KE, Williams DL, Kreiswirth BN, Musser JM. Characterization of rpsL and rrs mutations in streptomycin-resistant Mycobacterium tuberculosis isolates from diverse geographic localities. Antimicrob Agents Chemother. 1996;40:1024–6.
25. Tudó G, Rey E, Borrell S, Alcaide F, Codina G, Coll P, et al. Characterization of mutations in streptomycin-resistant Mycobacterium tuberculosis clinical isolates in the area of Barcelona. J Antimicrob Chemother. 2010;65:2341–6.
26. Springer B, Kidan YG, Prammananan T, Ellrott K, Böttger EC, Sander P. Mechanisms of streptomycin resistance: selection of mutations in the 16S rRNA gene conferring resistance. Antimicrob Agents Chemother. 2001;45:2877–84.
27. Wong SY, Lee JS, Kwak HK, Via LE, Boshoff HI, Barry CE. Mutations in gidB confer low-level streptomycin resistance in Mycobacterium tuberculosis. Antimicrob Agents Chemother. 2011;55:2515–22.
28. Jagielski T, Ignatowska H, Bakuła Z, Dziewit Ł, Napiórkowska A, Augustynowicz-Kopeć E, et al. Screening for streptomycin resistance-conferring mutations in Mycobacterium tuberculosis clinical isolates from Poland. PLoS One. 2014;9:e100078.
29. Safi H, Lingaraju S, Amin A, Kim S, Jones M, Holmes M, et al. Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-β-D-arabinose biosynthetic and utilization pathway genes. Nat Genet. 2013;45:1190–7.
30. Sreevatsan S, Stockbauer KE, Pan X, Kreiswirth BN, Moghazeh SL, Jacobs WR, et al. Ethambutol resistance in Mycobacterium tuberculosis: critical role of embB mutations. Antimicrob Agents Chemother. 1997;41:1677–81.
31. Newport MJ, Finan C. Genome-wide association studies and susceptibility to infectious diseases. Brief Funct Genomics. 2011;10:98–107.
32. Vincent V, Rigouts L, Nduwamahoro E, Holmes B, Cunningham J, Guillerm M, et al. The TDR Tuberculosis Strain Bank: a resource for basic science, tool development and diagnostic services. Int J Tuberc Lung Dis. 2012;16:24–31.
33. Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, et al. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. Nat Commun. 2014;5:4812.
34. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient control of population structure in model organism association mapping. Genetics. 2008;178:1709–23.
35. Cambau E, Viveiros M, Machado D, Raskine L, Ritter C, Tortoli E, et al. Revisiting susceptibility testing in MDR-TB by a standardized quantitative phenotypic assessment in a European multicentre study. J Antimicrob Chemother. 2015;70:686–96.
36. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nat Genet. 2013;45:1183–9.
37. Pires DE, Chen J, Blundell TL, Ascher DB. In silico functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. Sci Rep. 2016;6:19848. doi:10.1038/srep19848.
38. Usher JL, Ascher DB, Pires DE, Milan AM, Blundell TL, Ranganath LR. Analysis of HGD gene mutations in patients with alkaptonuria from the United Kingdom: identification of novel mutations. JIMD Rep. 2015;24:3–11.
39. Nemethova M, Radvanszky J, Kadasi L, Ascher DB, Pires DE, Blundell TL, et al. Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on "black bone disease" in Italy. Eur J Hum Genet. 2016;24:66–72.
40. Jafri M, Wake NC, Ascher DB, Pires DE, Gentle D, Morris MR, et al. Germline mutations in the CDKN2B tumor suppressor gene predispose to renal cell carcinoma. Cancer Discov. 2015;5:723–9.
41. van Soolingen D, Hermans PW, de Haas PE, Soll DR, van Embden JD. Occurrence and stability of insertion sequences in Mycobacterium tuberculosis complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. J Clin Microbiol. 1991;29:2578–86.
42. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
43. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30:2843–51.
44. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
45. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.
46. Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, et al. Fast computation and applications of genome mappability. PLoS One. 2012;7:e30377.
47. Stamatakis A, Hoover P, Rougemont J. A rapid bootstrap algorithm for the RAxML web servers. Syst Biol. 2008;57:758–71.
48. Coll F, Mallard K, Preston MD, Bentley S, Parkhill J, McNerney R, et al. SpolPred: rapid and accurate prediction of Mycobacterium tuberculosis spoligotypes from short genomic sequences. Bioinformatics. 2012;28:2991–3.
49. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38:904–9.
50. Reddy TB, Riley R, Wymore F, Montgomery P, DeCaprio D, Engels R, et al. TB database: an integrated platform for tuberculosis research. Nucleic Acids Res. 2009;37(Database issue):D499–508.
51. Lew JM, Kapopoulou A, Jones LM, Cole ST. TubercuList–10 years after. Tuberculosis (Edinb). 2011;91:1–7.
52. Alam MT, Petit RA, Crispell EK, Thornton TA, Conneely KN, Jiang Y, et al. Dissecting vancomycin-intermediate resistance in staphylococcus aureus using genome-wide association. Genome Biol Evol. 2014;6:1174–85.
53. Bertrand T, Eady NA, Jones JN, Jesmin, Nagy JM, Jamart-Grégoire B, et al. Crystal structure of Mycobacterium tuberculosis catalase-peroxidase. J Biol Chem. 2004;279:38991–9.
54. Velankar S, van Ginkel G, Alhroub Y, Battle GM, Berrisford JM, Conroy MJ, et al. PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. Nucleic Acids Res. 2016;44:D385–95.

55. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31:455–61.
56. Pires DE, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics. 2014;30:335–42.
57. Pires DE, Ascher DB, Blundell TL. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res. 2014;42(Web Server issue):W314–9.
58. Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZ, et al. Rapid whole genome sequencing of Mycobacterium tuberculosis isolates directly from clinical samples. J Clin Microbiol. 2015;53:2230–7.
59. Hoffman GE. Correcting for population structure and kinship using the linear mixed model: theory and extensions. PLoS One. 2013;8:e75707.
60. Xu M, Zhou YN, Goldstein BP, Jin DJ. Cross-resistance of Escherichia coli RNA polymerases conferring rifampin resistance to different antibiotics. J Bacteriol. 2005;187:2783–92.

Additional file 1: Table S1
The samples according to geographic location and phenotypic drug resistance

		Number of isolates belonging to lineage				**Number of isolates showing resistance to**								
Source	**Tot.**	**L1**	**L2**	**L3**	**L4**	**SM**	**INH**	**RMP**	**EMB**	**OFL**	**KAN**	**CAP**	**Et**	**PZA**
Asia														
Bangladesh	8	4	1	1	2	2	4	1	4	1	-	-	1	-
China (Tibet)	1	1	-	-	-	1	1	1	-	-	-	-	-	-
Nepal	4	1	2	-	1	4	3	2	1	1	-	-	-	2
Pakistan	1	-	-	1	-	-	-	-	-	-	-	-	-	-
Philippines	4	4	-	-	-	1	2	2	1	-	-	-	2	-
Sth Korea	39	-	23	1	15	17	26	17	15	-	-	1	1	3
Thailand	1	-	1	-	-	-	-	-	-	1	1	1	-	-
Africa														
Cameroon	1	-	-	-	1	1	-	-	-	-	-	-	-	-
CAR	1	-	-	-	1	-	-	-	-	1	-	-	-	-
Guinea	1	-	-	-	1	-	1	-	-	-	-	-	-	-
Guinea Eq.	1	-	-	-	1	-	1	1	-	-	-	-	1	-
Morocco	4	-	-	-	4	2	3	1	1	-	-	-	-	-
Niger	1	-	-	-	1	-	-	-	-	-	1	1	-	-
Nigeria	2	-	-	-	2	-	1	1	-	1	-	-	1	-
RDC	4	-	-	-	4	-	-	-	-	-	1	1	-	-
Rwanda	15	-	-	-	15	4	15	15	10	-	-	-	1	-

Europe														
Germany	12	-	1	1	10	3	1	1	1	1	1	1	-	-
Kazakhstan	1	-	-	-	1	-	-	1	-	-	-	-	-	-
Portugal	1	-	-	-	1	1	1	1	-	-	-	-	-	-
Spain	2	-	-	-	2	1	-	1	-	-	-	-	-	-
South America														
Brazil	7	-	-	-	7	2	4	4	2	1	-	-	1	1
Colombia	1	-	-	-	1	-	1	1	-	-	-	-	-	-
Peru	31	-	5	-	26	9	9	11	6	1	-	1	1	-
Rep. Domin.	1	-	-	-	1	-	-	-	-	-	-	-	-	-
Overall	**144**	**10**	**33**	**4**	**97**	**48**	**73**	**61**	**41**	**8**	**4**	**6**	**9**	**6**

CAR Central African Republic; DRC Democratic Republic of Congo, L1-L4 lineages 1 to 4, (first line drugs) RMP = Rifampicin, INH = Isoniazid, SM = Streptomycin, EMB = Ethambutol; (second line drugs) OFL = Ofloxacin , KAN = kanamycin, CAP = capreomycin, Et = ethionamide, P =Para-aminosalisylic acid.

Additional file 1: Table S2
The isolate ENA accession numbers and MIC values

ENA Accession	TDR Accession	RMP MIC	INH MIC	SM MIC	EMB MIC
ERR1213824	TB-TDR-0070	>120	3.2	4	8
ERR1213825	TB-TDR-0073	>120	0.8	≤1	4
ERR1213826	TB-TDR-0074	80	3.2	2	≤1
ERR1213827	TB-TDR-0077	30	0.2	2	≤1
ERR1213828	TB-TDR-0078	≤10	0.2	≤1	2
ERR1213829	TB-TDR-0079	80	0.2	4	≤1
ERR1213830	TB-TDR-0080	≤10	0.2	≤1	≤1
ERR1213831	TB-TDR-0081	≤10	0.2	≤1	2
ERR1213832	TB-TDR-0082	≤10	3.2	≤1	2
ERR1213833	TB-TDR-0083	≤10	3.2	≤1	≤1
ERR1213834	TB-TDR-0084	≤10	3.2	≤1	≤1
ERR1213835	TB-TDR-0085	≤10	3.2	8	4
ERR1213836	TB-TDR-0086	>120	0.8	≤1	2
ERR1213837	TB-TDR-0087	>120	3.2	8	4
ERR1213838	TB-TDR-0088	>120	0.2	≤1	≤1
ERR1213839	TB-TDR-0089	>120	0.8	≤1	8
ERR1213840	TB-TDR-0090	>120	0.2	≤1	2
ERR1213841	TB-TDR-0091	20	0.2	2	2
ERR1213842	TB-TDR-0092	≤10	0.2	2	4
ERR1213843	TB-TDR-0093	≤10	>3.2	8	4
ERR1213844	TB-TDR-0094	≤10	0.2	≤1	8
ERR1213845	TB-TDR-0095	≤10	0.8	>16	≤1
ERR1213846	TB-TDR-0096	≤10	≤0.05	>16	≤1
ERR1213847	TB-TDR-0097	30	0.2	>16	≤1
ERR1213848	TB-TDR-0098	40	0.2	16	2
ERR1213849	TB-TDR-0099	>120	>3.2	4	>8
ERR1213850	TB-TDR-0101	80	3.2	2	4
ERR1213851	TB-TDR-0102	>120	3.2	≤1	2
ERR1213852	TB-TDR-0104	≤10	>3.2	≤1	4
ERR1213853	TB-TDR-0106	>120	3.2	≤1	2
ERR1213854	TB-TDR-0108	≤10	0.8	8	≤1
ERR1213855	TB-TDR-0109	≤10	3.2	>16	2
ERR1213856	TB-TDR-0110	≤10	>3.2	>16	2
ERR1213857	TB-TDR-0112	>120	0.8	>16	≤1
ERR1213858	TB-TDR-0113	120	>3.2	>16	4
ERR1213859	TB-TDR-0116	80	>3.2	8	>8
ERR1213860	TB-TDR-0117	>120	>3.2	≤1	2
ERR1213861	TB-TDR-0119	>120	3.2	4	≤1
ERR1213862	TB-TDR-0120	30	>3.2	16	4
ERR1213863	TB-TDR-0122	>120	3.2	>16	4
ERR1213864	TB-TDR-0123	20	>3.2	>16	8
ERR1213865	TB-TDR-0124	>120	>3.2	>16	4

ERR1213866	TB-TDR-0125	>120	0.2	≤1	2
ERR1213867	TB-TDR-0126	≤10	0.2	2	≤1
ERR1213868	TB-TDR-0129	>120	1.6	2	≤1
ERR1213869	TB-TDR-0130	≤10	≤0.05	2	≤1
ERR1213870	TB-TDR-0131	>120	>3.2	>16	≤1
ERR1213871	TB-TDR-0132	≤10	>3.2	4	4
ERR1213872	TB-TDR-0133	>120	1.6	8	4
ERR1213873	TB-TDR-0134	>120	>3.2	4	≤1
ERR1213874	TB-TDR-0135	>120	>3.2	≤1	>8
ERR1213875	TB-TDR-0136	≤10	≤0.05	2	≤1
ERR1213876	TB-TDR-0137	≤10	≤0.05	2	≤1
ERR1213877	TB-TDR-0138	≤10	≤0.05	2	2
ERR1213878	TB-TDR-0139	≤10	0.2	2	2
ERR1213879	TB-TDR-0140	≤10	0.2	≤1	≤1
ERR1213880	TB-TDR-0141	>120	0.2	>16	≤1
ERR1213881	TB-TDR-0142	20	0.2	2	2
ERR1213882	TB-TDR-0143	≤10	0.2	16	2
ERR1213883	TB-TDR-0144	≤10	>3.2	>16	4
ERR1213884	TB-TDR-0146	40	0.2	8	≤1
ERR1213885	TB-TDR-0147	20	>3.2	>16	≤1
ERR1213886	TB-TDR-0148	>120	>3.2	>16	2
ERR1213887	TB-TDR-0149	>120	>3.2	>16	4
ERR1213888	TB-TDR-0150	>120	1.6	2	≤1
ERR1213889	TB-TDR-0152	120	3.2	8	8
ERR1213890	TB-TDR-0153	80	3.2	4	2
ERR1213891	TB-TDR-0155	120	>3.2	2	2
ERR1213892	TB-TDR-0156	≤10	0.2	2	4
ERR1213893	TB-TDR-0157	>120	0.2	>16	2
ERR1213894	TB-TDR-0158	≤10	≤0.05	2	≤1
ERR1213895	TB-TDR-0159	≤10	≤0.05	2	≤1
ERR1213896	TB-TDR-0160	80	0.2	8	≤1
ERR1213897	TB-TDR-0161	≤10	≤0.05	≤1	≤1
ERR1213898	TB-TDR-0163	≤10	≤0.05	≤1	≤1
ERR1213899	TB-TDR-0164	≤10	≤0.05	8	≤1
ERR1213900	TB-TDR-0165	>120	0.2	2	≤1
ERR1213901	TB-TDR-0166	>120	3.2	>16	2
ERR1213902	TB-TDR-0167	>120	3.2	16	4
ERR1213903	TB-TDR-0169	≤10	≤0.05	>16	≤1
ERR1213904	TB-TDR-0170	>120	>3.2	2	4
ERR1213905	TB-TDR-0171	>120	0.2	≤1	≤1
ERR1213906	TB-TDR-0172	20	>3.2	2	≤1
ERR1213907	TB-TDR-0173	≤10	≤0.05	≤1	≤1
ERR1213908	TB-TDR-0174	≤10	≤0.05	2	≤1
ERR1213909	TB-TDR-0175	>120	0.8	≤1	4
ERR1213910	TB-TDR-0176	>120	1.6	>16	2
ERR1213911	TB-TDR-0177	20	0.2	2	≤1

ERR1213912	TB-TDR-0178	≤10	0.8	2	≤1
ERR1213913	TB-TDR-0180	≤10	0.2	2	≤1
ERR1213914	TB-TDR-0181	≤10	≤0.05	8	2
ERR1213915	TB-TDR-0182	20	≤0.05	>16	≤1
ERR1213916	TB-TDR-0183	>120	0.8	2	≤1
ERR1213917	TB-TDR-0184	≤10	≤0.05	2	≤1
ERR1213918	TB-TDR-0185	>120	0.8	2	≤1
ERR1213919	TB-TDR-0186	>120	>3.2	≤1	≤1
ERR1213920	TB-TDR-0187	≤10	>3.2	8	2
ERR1213921	TB-TDR-0189	>120	>3.2	8	≤1
ERR1213922	TB-TDR-0190	>120	≤0.05	>16	≤1
ERR1213923	TB-TDR-0191	>120	>3.2	8	8
ERR1213924	TB-TDR-0193	>120	>3.2	>16	2
ERR1213925	TB-TDR-0194	20	0.2	2	≤1
ERR1213926	TB-TDR-0195	20	3.2	2	8
ERR1213927	TB-TDR-0197	20	0.2	≤1	2
ERR1213928	TB-TDR-0198	>120	3.2	8	4
ERR1213929	TB-TDR-0199	≤10	0.2	≤1	2
ERR1213930	TB-TDR-0200	≤10	≤0.05	≤1	≤1
ERR1213931	TB-TDR-0201	>120	>3.2	16	≤1
ERR1213932	TB-TDR-0202	20	≤0.05	4	2
ERR1213933	TB-TDR-0203	≤10	≤0.05	2	≤1
ERR1213934	TB-TDR-0204	20	≤0.05	4	≤1
ERR1213935	TB-TDR-0207	≤10	≤0.05	2	≤1
ERR1213936	TB-TDR-0208	20	0.2	≤1	≤1
ERR1213937	TB-TDR-0209	20	0.2	2	≤1
ERR1213938	TB-TDR-0210	20	≤0.05	≤1	≤1
ERR1213939	TB-TDR-0213	20	0.8	2	≤1
ERR1213940	TB-TDR-0214	≤10	≤0.05	2	≤1
ERR1213941	TB-TDR-0016	≤10	0.2	≤1	2
ERR1213942	TB-TDR-0017	≤10	0.2	2	≤1
ERR1213943	TB-TDR-0018	>120	0.2	≤1	≤1
ERR1213944	TB-TDR-0022	20	≤0.05	2	≤1
ERR1213945	TB-TDR-0038	≤10	>3.2	>16	≤1
ERR1213946	TB-TDR-0041	≤10	≤0.05	2	≤1
ERR1213947	TB-TDR-0042	≤10	>3.2	≤1	>8
ERR1213948	TB-TDR-0043	20	3.2	≤1	8
ERR1213949	TB-TDR-0045	≤10	0.2	2	4
ERR1213950	TB-TDR-0007	>120	>3.2	>16	4

RMP rifampicin, INH isoniazid, SM streptomycin, EMB ethambutol

Additional file 1: Table S3
Drug susceptibility profiles for rifampicin, isoniazid, streptomycin and ethambutol

No. samples	Rifampicin	Isoniazid	Streptomycin	Ethambutol
12 (9.4%)	***R***	***R***	***R***	***R***
8 (6.3%)	***R***	***R***	***R***	*S*
8 (6.3%)	***R***	***R***	*S*	***R***
14 (11.0%)	***R***	***R***	*S*	*S*
4 (3.1%)	**R**	S	**R**	S
7 (5.5%)	**R**	S	S	S
5 (3.4%)	S	**R**	**R**	**R**
7 (5.5%)	S	**R**	**R**	S
5 (3.4%)	S	**R**	S	**R**
6 (4.7%)	S	**R**	S	S
9 (7.1%)	S	S	**R**	**S**
4 (3.1%)	S	S	S	**R**
38 (29.9%)	S	S	S	S

R = resistance, S = sensitive; 13 different profiles were identified across 127 independent samples; Multi-drug resistant in italics

Additional file 1: Table S4
Combinations of mutations and their frequency (N) in drug resistance candidate genes

a) Rifampicin

Mutation observed in *rpoB* codons										*rpoC*	N	MIC (µg/ml)		
	1	2	4	4	4	4	4	6						
4	7	5	0	3	4	5	9	9						
5	0	0	0	5	5	0	1	2	rare			Mean	Min	Max
											70	15.3	10.0	80.0
									*		5	68.0	10.0	120.0
								*			1	10.0	10.0	10.0
							*				1	80.0	80.0	80.0
						*					21	114.3	80.0	120.0
						*				*	1	120.0	120.0	120.0
						*			*		2	120.0	120.0	120.0
						*			*	*	1	120.0	120.0	120.0
						*	*				1	120.0	120.0	120.0
						**					1	120.0	120.0	120.0
					*						9	120.0	120.0	120.0
				*							4	120.0	120.0	120.0
				*				*			1	120.0	120.0	120.0
			*			*					2	120.0	120.0	120.0
		*									2	10.0	10.0	10.0
	*										1	120.0	120.0	120.0
	*								*		1	120.0	120.0	120.0
*						*					3	120.0	120.0	120.0

b) Isoniazid

			inhA		MIC (µg/ml)		
katG codons			prom.	N			
315	436	rare			Mean	Min	Max
				46	0.3	0.05	3.2
		*		2	0.2	0.2	0.2
		***		1	3.2	3.2	3.2
			*	8	1.7	0.8	3.2
		*	*	1	1.6	1.6	1.6
*				23	2.9	0.2	3.2
*		*		1	3.2	3.2	3.2
*			*	2	3.2	3.2	3.2
	*			18	0.2	0.05	0.8
	*		*	4	1.0	0.8	1.6
	*	*	*	1	3.2	3.2	3.2
*	*			18	3.2	3.2	3.2
*	*	*		1	3.2	3.2	3.2
*	*		*	1	3.2	3.2	3.2

c) Streptomycin

rpsL codons		*rrs*	N	MIC (µg/ml)		
43	88			Mean	Min	Max
			99	3.8	1	16
		*	13	10.0.4	1	16
	*		4	16.0	16	16
*			11	16.0	16	16

d) Ethambutol

embB codons									*ubiA*	*embA*	N	MIC (µg/ml)		
297	306	319	354	378	406	497	1024	rare				Mean	Min	Max
											70	1.2	1	2
										**	1	4.0	4	4
									*		2	1.5	1	2
									**		1	1.0	1	1
								*			3	1.7	1	2
								**			1	2.0	2	2
							*				1	4.0	4	4
						*					5	4.8	4	8
						*			*		1	4.0	4	4
					*						6	4.5	1	8
					*				*		4	2.8	1	4
					*		*				1	4.0	4	4
					*		*	*			1	4.0	4	4
				*					**		3	1.0	1	1
				*		*		*	**		1	2.0	2	2
			*							*	1	4.0	4	4
			*	*					**		1	4.0	4	4
			*	*				*	**		1	4.0	4	4
		*									3	4.7	2	8
	*										11	4.1	1	8
	*									*	1	4.0	4	4
	*								*		1	8.0	8	8
	*							*			1	2.0	2	2
	*			*					**		2	8.0	8	8
	*		*								1	8.0	8	8
*											2	2.5	1	4
*										*	1	8.0	8	8

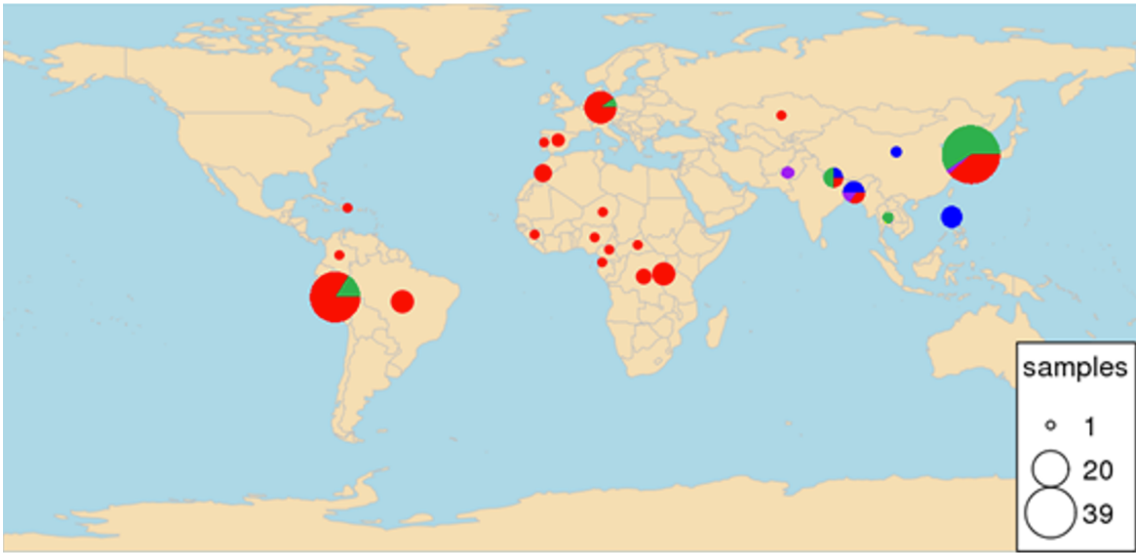
* single mulation, ** double mutations, *** triple mutations; SNP mutations in a single sample have been aggregated into a “rare” column.

Additional file 1: Table S5
Predicted effects of mutations

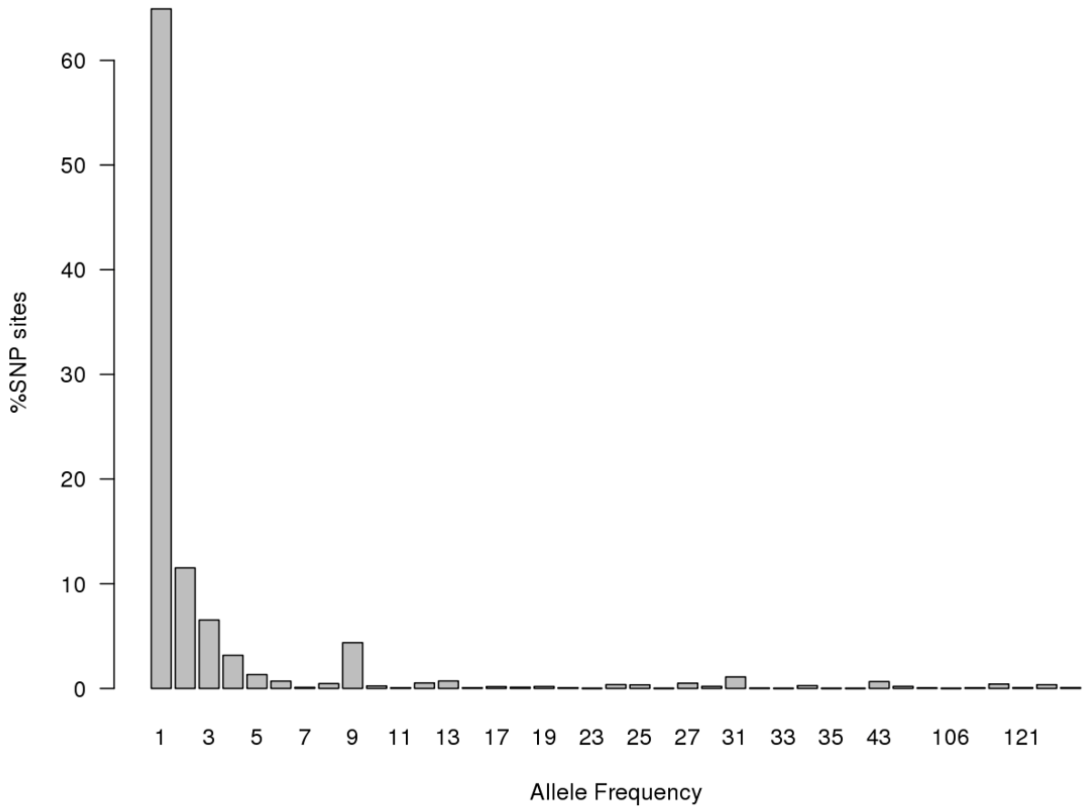
Gene	Mutation	Distance to interface (Å)	Distance to ligand (Å)	DUET (ΔΔG kcal/mol)	mCSM-Stability (ΔΔG kcal/mol)	SDM (ΔΔG kcal/mol)
rpoB	T400A		42.914	0.031	-0.326	2.480
	D435V		3.094	0.336	0.356	1.860
	H445D		4.015	-2.084	-1.971	-1.730
	H445Y		4.015	-0.214	-0.171	-0.310
	H445R		4.015	-1.958	-1.857	-1.950
	S450W		5.773	-0.756	-0.840	2.330
	S450L		5.773	0.102	-0.126	2.820
	I491V		2.908	-1.221	-1.274	-0.830
	I491F		2.908	-1.529	-1.416	-0.760
katG	S315N	14.940	2.149	-0.100	-0.184	2.149
	S315T	14.940	2.149	-0.243	-0.330	2.149

Additional file 2: Figure S1

The global distribution of geographic origin and lineage of the isolates. Lineages one to four are represented by blue, green, purple, and red, respectively



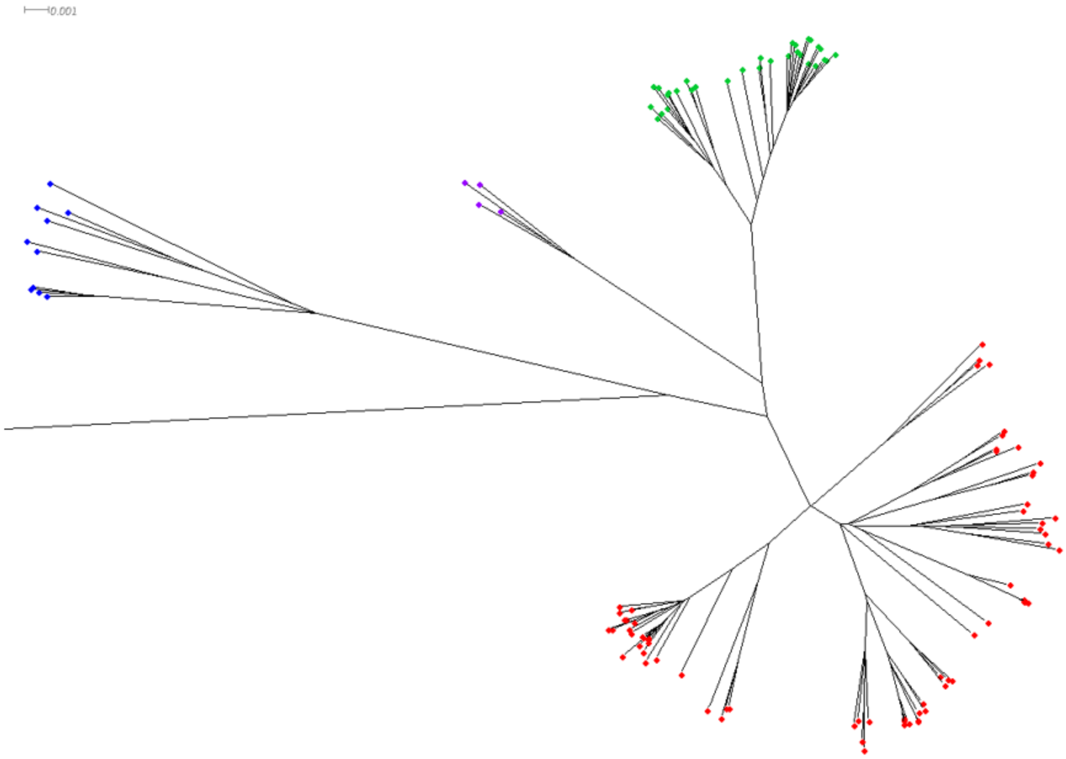
Additional File 3: Figure S2
SNP allele frequency spectrum. A large number of rare variants are observed. Peaks with higher allele frequency reflect the presence of lineage and sub-lineage specific SNPs.



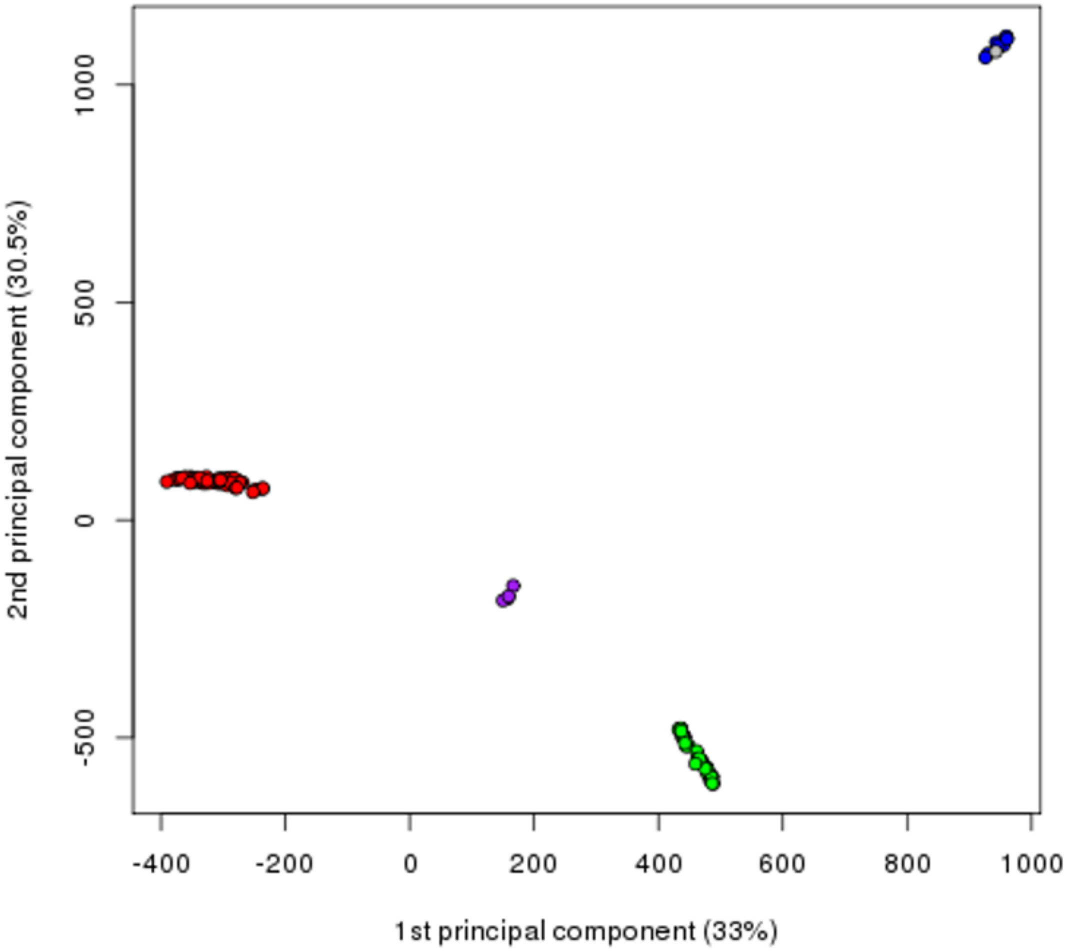
Additional File 4: Figure S3

Population structure analysis of the 144 isolates show clustering by lineage (Lineages one to four are represented by blue, green, purple, and red points, respectively). (a) A phylogenetic tree rooted with *M. canetti*. (b) First two principal components represent 33 % and 30.5 % of the variation explained between isolates, respectively.

a)



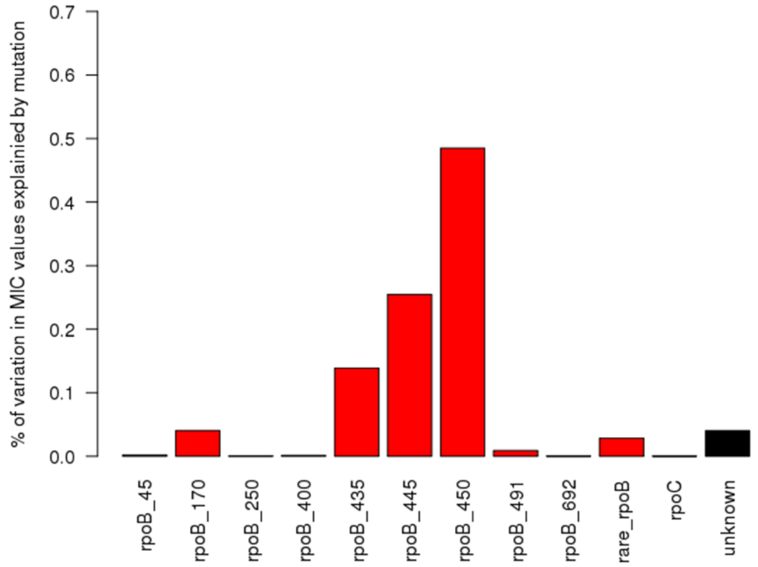
b)


2nd principal component (30.5%)
1000
500
0
-500
-400
-200
0
200
400
600
800
1000
1st principal component (33%)

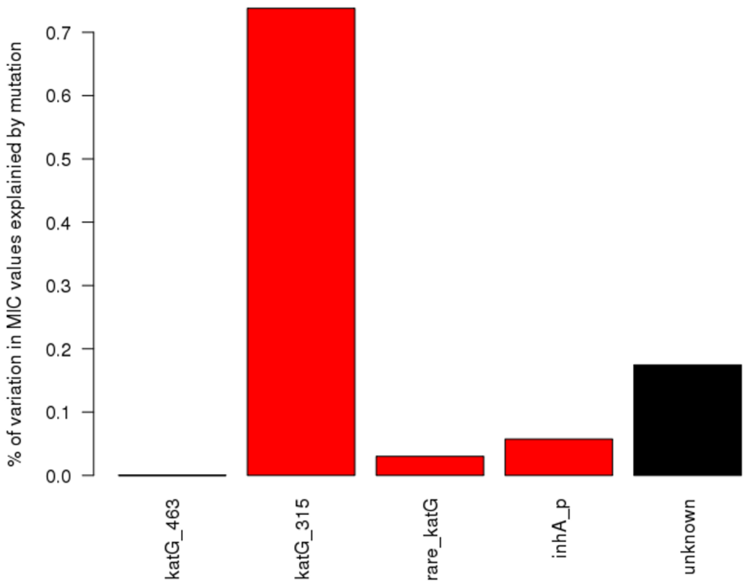
Additional file 6: Figure S5

Percentage of the variation in MIC values explained by each mutated codon in candidate genes. Bars in red represent significant independent associations with increased MIC (p < 0.05). a) Rifampicin. b) Isoniazid. c) Streptomycin. d) Ethambutol.

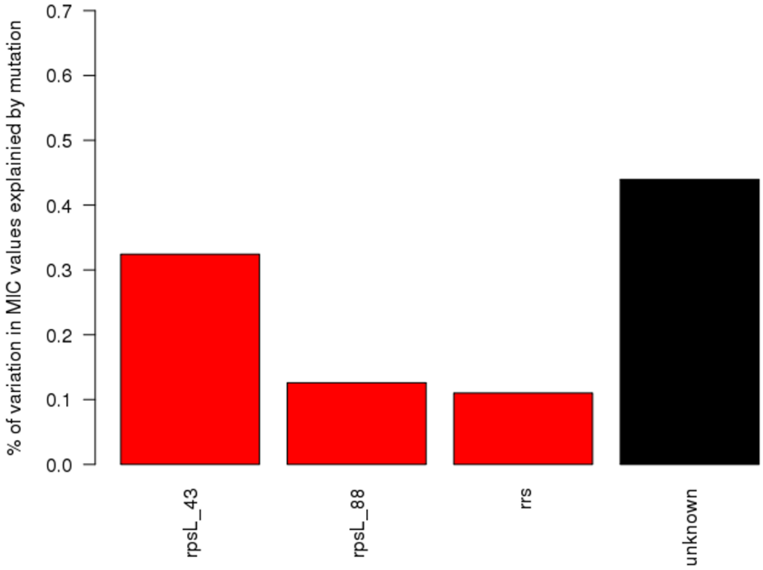
a)



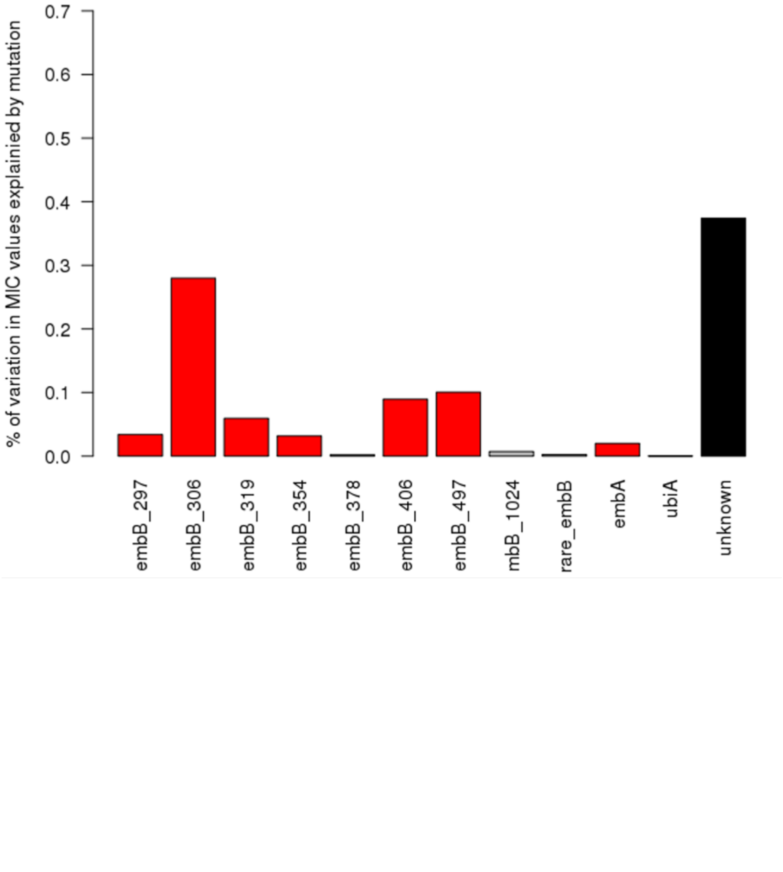
b)



c)

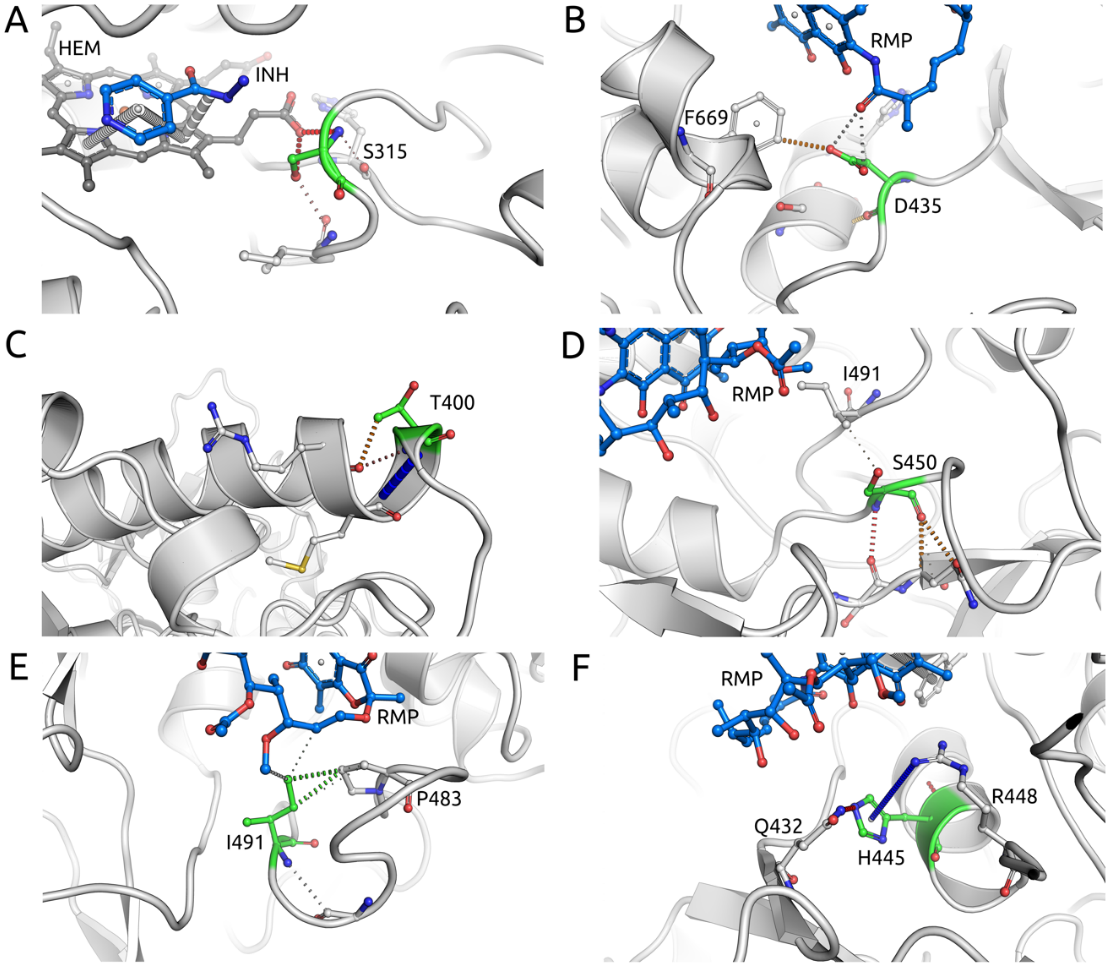

% of variation in MIC values explainied by mutation
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
rpsL_43
rpsL_88
rrs
unknown

d)


% of variation in MIC values explainied by mutation
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
embB_297
embB_306
embB_319
embB_354
embB_378
embB_406
embB_497
mbB_1024
rare_embB
embA
ubiA
unknown

Additional file 7: Figure S6

Molecular interactions established by wild-type residues in katG and rpoB residues. (A) The interactions established by Ser315 in katG. Given the proximity of the residue to the ligands INH and HEM, mutations to Asn and Thr, with slightly larger side chains, would potentially cause steric clashes. (B) The interactions of Asp435 in rpoB. It directly interacts with RMP via polar interactions that would be disrupted by mutations to Val. (C) Thr400 in rpoB is at the end of an alpha helix establishing intra molecular interactions. Giving its distance to RMP, it would be expected that its mutation to Ala would be a lower impact, which would arise from alosteric changes. (D) Ser450 establishes strong intra molecular interactions in the RMP binding site. Mutations to larger residues (Trp and Leu) could disrupt the packing of the region and therefore binding. (E). Ile491 performs hydrophobic interactions with RMP and its neighbouring residues. Mutations to Phe or Val would compromise packing, either inducing steric clashes or compromising packing. (F). His445 performs strong intra molecular interactions, including a donor-pi (blue dashes) and hydrogen bond (red dashes). Mutations to residues Asp, Tyr or Arg would imply in the loss of the pi interaction as well as potential introduction of steric clashes.



RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Jody Phelan
Principal Supervisor	Taane Clark
Thesis Title	A Bioinformatic analysis of M. tuberculosis and host genomic data

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Nature Genetics
Please list the paper's authors in the intended authorship order:	Francesc Coll, Jody Phelan, Grant A. Hill-Cawthorne, Mridul B Nair, Kim Mallard, Shahjahan Ali, Abdallah M Abdallah, Saad Alghamdi, Mona Alsomali, Abdallah O Ahmed, Stephanie Portelli , Yaa Oppong, Adriana Alves, Theolis Barbosa Bessa, Susana Campino, Maxine Caws, Anirvan Chatterjee, Amelia C Crampin, Keertan Dheda, Nicholas Furnham , Judith R Glynn, Louis Grandjean, Dang Thi Minh Ha, Rumina Hasan, Zahra Hasan, Martin L Hibberd, Moses Joloba, Edward C. Jones-López, Tomoshige Matsumoto, Anabela Miranda, David Moore, Nora Mocillo, Stefan Panaiotov, Julian Parkhill, Carlos Penha, João Perdigão, Isabel Portugal, Zineb Rchiad, Jaime Robledo, Patricia Sheen, Nashwa Talaat Shesha, Frik A Sirgel, Christophe Sola, Erivelton de Oliveira Sousa, Elizabeth M Streicher, Paul Van Helden , Miguel Viveiros, Robert M Warren, Ruth McNerney , Arnab Pain, Taane G Clark
Stage of publication	**In press**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I took over this project from the previous PhD student at the laboratory of Taane Clark (Francesc Coll). I downloaded additional data from the ENA along with any available meta data. I then curated and analysed the final NGS dataset and wrote custom scripts in python and R to perform all analyses, except the protein modelling work, but including and not limited to sequence QC, mapping, variant calling and QC, phylogenetic reconstruction, optimisation of GWAS parameters and curation of the final results set. I generated all figures using R except supplementary Figure 3. I co-wrote a draft of the first paper with Taane Clark and circulated this to other co-authors. After incorporating all the comments, I was actively involved in the submission process to Nature Genetics and performed all subsequent analyses which was required during the revision stage.

Student Signature: ______________________________ **Date:** ______________

Supervisor Signature: ______________________________ **Date:** ______________

Chapter 4

The *Mycobacterium tuberculosis* resistome from a genome-wide analysis of multi- and extensively drug-resistant tuberculosis

The Mycobacterium tuberculosis resistome from a genome-wide analysis of multi- and extensively drug-resistant tuberculosis

Francesc Coll[1,*], Jody Phelan[1*], Grant A. Hill-Cawthorne[2,3**], Mridul B Nair[2,**], Kim Mallard[1], Shahjahan Ali[2], Abdallah M Abdallah[2], Saad Alghamdi[4], Mona Alsomali[2], Abdallah O Ahmed[5], Stephanie Portelli [1], Yaa Oppong[1], Adriana Alves[6], Theolis Barbosa Bessa[7], Susana Campino[1], Maxine Caws[8,9], Anirvan Chatterjee[10], Amelia C Crampin[11,12], Keertan Dheda[13], Nicholas Furnham [1], Judith R Glynn[11,12], Louis Grandjean[14], Dang Thi Minh Ha[9], Rumina Hasan[15], Zahra Hasan[15], Martin L Hibberd[1], Moses Joloba[16], Edward C. Jones-López[17], Tomoshige Matsumoto[18], Anabela Miranda[6], David J Moore[1,14], Nora Mocillo[19], Stefan Panaiotov[20], Julian Parkhill[21], Carlos Penha[22], João Perdigão[23], Isabel Portugal[23], Zineb Rchiad[4], Jaime Robledo[24], Patricia Sheen[13], Nashwa Talaat Shesha[25], Frik A Sirgel[26], Christophe Sola[27], Erivelton de Oliveira Sousa[28], Elizabeth M Streicher[26], Paul Van Helden [26], Miguel Viveiros[29], Robert M Warren[26], Ruth McNerney [1,13,***], Arnab Pain[2,30,***], Taane G Clark[1,11,***]

1 Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT, United Kingdom

2 Pathogen Genomics Laboratory, BESE Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia

3 Sydney Emerging Infections and Biosecurity Institute and School of Public Health, Sydney Medical School, University of Sydney, NSW 2006, Australia

4 Laboratory Medicine Department, Faculty of Applied Medical Sciences, Umm Al-Qura University, Kingdom of Saudi Arabia

5 Department of Microbiology, Faculty of Medicine, Umm Al-Qura University, Makkah, Saudi Arabia

6 National Mycobacterium Reference Laboratory, Porto, Portugal

7 Centro de Pesquisas Goncalo Moniz, Fundacao Oswaldo Cruz Bahia R. Waldemar Falcao 121 Candeal 40296-710 Salvador Bahia Brazil

8 Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, L3 5QA, United Kingdom

9 Pham Ngoc Thach Hospital for TB and Lung Diseases, Hung Vuong, Ho Chi Minh City, Vietnam

10 The Foundation for Medical Research, 84-A, R. G. Thadani Marg, Worli, Mumbai 400018, India

11 Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT, United Kingdom

12 Karonga Prevention Study, Malawi

13 Lung Infection and Immunity Unit, UCT Lung Institute, University of Cape Town, Groote Schuur Hospital, Observatory, 7925, Cape Town, South Africa.

14 Laboratorio de Enfermedades Infecciosas, Laboratorios de Investigación y Desarrollo, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Lima, Peru

15 Department of Pathology and Laboratory Medicine, The Aga Khan University, Stadium Road, P.O. Box 3500, Karachi 74800, Pakistan

16 Department of Medical Microbiology, Makerere University College of Health Sciences, Kampala, Uganda

17 Section of Infectious Diseases, Department of Medicine, Boston Medical Center and Boston University School of Medicine, Boston, Massachusetts, USA

18 Osaka Anti-Tuberculosis Association Osaka Hospital, Osaka, Japan

19 Reference Laboratory of Tuberculosis Control, Buenos Aires, Argentina

20 National Center of Infectious and Parasitic Diseases, 1504 Sofia, Bulgaria

21 Wellcome Trust Sanger Institute, Hinxton, United Kingdom

22 Instituto Gulbenkian de Ciência, Lisbon, Portugal

23 iMed.ULisboa - Research Institute for Medicines, Faculdade de Farmácia, Universidade de Lisboa, Portugal

24 Corporación para Investigaciones Biológicas, Universidad Pontificia Bolivariana, Medellín, Colombia

25 Regional Laboratory Directorate of Health Affairs, Makkah, Kingdom of Saudi Arabia.

26 Division of Molecular Biology and Human Genetics, SAMRC Centre for Tuberculosis Research, DST/NRF Centre of Excellence for Biomedical Tuberculosis Research, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, South Africa

27 Institute for Integrative Cell Biology, CEA-CNRS-Université Paris-Saclay, Orsay, France

28 Centro de Pesquisas Goncalo Moniz, Fundacao Oswaldo Cruz Bahia R. Waldemar Falcao 121 Candeal 40296-710 Salvador Bahia Brazil and Laboratorio Central de Saude Publica Prof. Goncalo Moniz Rua Waldemar Falcao, 123 Horto Florestal 40295-010 Salvador Bahia Brazil

29 Unidade de Microbiologia Médica, Global Health and Tropical Medicine, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, UNL, Lisboa, Portugal

30 Global Station for Zoonosis Control, Global Institution for Collaborative Research and Education (GI-CoRE), Hokkaido University, N20 W10 Kita-ku, Sapporo, 001-0020 Japan

[*] Joint first authors, contributed equally.

[**] Contributed equally.

[***] Corresponding authors: Taane Clark (e-mail: taane.clark@lshtm.ac.uk) or Arnab Pain (e-mail: arnab.pain@kaust.edu.sa), or Ruth McNerney (ruth.mcnerney@uct.ac.za)

ABSTRACT

To further characterize the genetic determinants of resistance to anti-tuberculosis drugs we performed a genome-wide association study (GWAS) of 6,465 *Mycobacterium tuberculosis* clinical isolates from more than 30 countries. A GWAS approach within a mixed regression framework was followed by a phylogenetic-based test for independent mutations. In addition to novel mutations associated with resistance to cycloserine, ethionamide and p-aminosalicylic acid our analysis indicates a more extensive role for small insertions and deletions and large deletions than previously recognised, particularly for ethionamide, pyrazinamide, capreomycin, cycloserine and para-aminosalicylic acid. Findings also suggest the involvement of efflux pumps (*drrA, Rv2688c*) in the emergence of resistance. Odds ratios for mutations in candidate loci were found to reflect levels of resistance reported from phenotypic testing. Findings from this study, the most comprehensive yet reported, will inform the design of new diagnostic tests and expedite the investigation of resistance and compensatory epistatic mechanisms.

KEY WORDS: *Mycobacterium tuberculosis***,** tuberculosis, GWAS, drug resistance, MDR-TB, XDR-TB

The emergence and spread of *Mycobacterium tuberculosis* (*Mtb*) resistant to multiple anti-tuberculous drugs is of global concern. Programmatically incurable tuberculosis (TB), where effective treatment regimens cannot be provided due to resistance to the available drugs is a growing problem[1]**.** In high burden countries such patients may be discharged to home care, increasing the risk of community based transmission of incurable drug resistant disease[2]. Resistance to rifampicin and isoniazid is classed as multidrug-resistant tuberculosis (MDR-TB), further resistance to the fluoroquinolones and any of the injectable drugs (amikacin, kanamycin or capreomycin) used to treat MDR-TB is termed extensively drug-resistant (XDR-TB). Strains resistant to the remaining drugs, are referred to as *XXDR-TB or* totally drug-resistant (TDR-TB), however formal definitions for post XDR-TB resistance have yet to be agreed by the World Health Organisation (WHO)[1]. Treatment for patients with drug resistant tuberculosis is prolonged, expensive and outcomes are poor[3]. The drugs used are toxic and poorly tolerated, adverse events are common and may be severe and irreversible[4]. Inadequate treatment also risks amplification of resistance to further drugs and may prolong opportunities for transmission[5]**.**

Mtb has a clonal genome (size 4.4Mb) with a low mutation rate and no evidence of between-strain recombination or horizontal gene transfer[6]. The *Mtb* complex comprises seven lineages, of which four are predominant in humans: Lineage 1, Indo-Oceanic (e.g. East-African-Indian (EAI) spoligotype families); Lineage 2, East-Asian (e.g. W/Beijing spoligotype families); Lineage 3, East-African-Indian (e.g. Central-Asian-Strain (e.g. CAS-DELHI) spoligotype families) and Lineage 4, Euro-American (e.g. *Latin American-Mediterranean (LAM),* Haarlem and the "ill-defined" T spoligotype families)*[6].*

Resistance in *Mtb* is mainly conferred by nucleotide variations (single nucleotide polymorphisms, insertions and deletions (indels)) in genes coding for drug-targets or -converting enzymes. Changes in efflux pump regulation may have an impact on the emergence of resistance[7] and putative compensatory mechanisms to overcome fitness impairment coincidental with the acquisition of resistance have been described for some drugs[8]**.** Detection of resistance conferring mutations offers a means of rapidly identifying resistance to anti-tuberculosis drugs[9] but, with the exception of rifampicin, current molecular tests for resistance lack high levels of sensitivity[9]**.** To improve knowledge of genetic determinants of drug resistance we undertook whole genome analysis of a large collection (n=6,465) of clinical isolates from more than 30 geographic locations, representing the four major *Mtb* lineages **(Figure 1, Supplementary table 1)**. We adopted a GWAS approach to identify nucleotide variation and loci underlying drug resistance as successfully applied in *Mtb*[10–12] and other bacteria[13,14]. A total of 14 drugs with available phenotypic data on drug susceptibility testing were investigated **(Supplementary table 2)**. Phenotypic drug susceptibility data was not available for each of the 14 drugs for every isolate and sample sizes ranged from over 6,000 for the most commonly tested first line drugs (isoniazid and rifampicin) to 255 and 248 for p-aminosalicylic acid and cycloserine, respectively, which are used to treat patients with XDR-TB. Here, we present findings from the most comprehensive study yet undertaken of the genetic determinants of resistance to anti-tuberculosis drugs or the *Mtb* resistome.

RESULTS

Genetic polymorphisms, population structure and drug resistance

High quality genome-wide SNPs (102,160), indels (11,122), and large deletions (284) were identified across all samples (n=6,465). The majority of SNPs (93.1%) had rare minor alleles (allele frequency <1%) (**Supplementary Figure 1**) and 23,216 SNPs (8.9%) were very rare (minor allele frequency <= 0.3%). Similarly, small indels were rare (96.6% had frequency <1%), and ranged in size from 1 to 45bp. The majority (82.7%, 7788/9421) in protein coding genes resulted in frame-shifts, leading to premature stop signals in the coding mRNAs. A phylogenetic tree and principal component analysis constructed using all genome-wide SNPs and small indels revealed the expected clustering by lineage (**Figure 2, Supplementary Figure 2**).

Phenotypic analysis of susceptibility to anti-tuberculosis drugs found 27.7% of isolates were resistant to at least one drug, 12.9% were categorised as MDR-TB and 4.5% as XDR-TB (**Supplementary table 2, Figure 2**). Fourteen drugs were included in the genome-wide analysis: isoniazid (INH), rifampicin (RIF), ethionamide (ETH), pyrazinamide (PZA), ethambutol (EMB), streptomycin (STM), amikacin (AMK), capreomycin (CAP), kanamycin (KAN), ciprofloxacin (CIP), ofloxacin (OFL), moxifloxacin (MOX), cycloserine (CYS) and para-aminosalicylic acid (PAS). Drug family groups including the second-line injectable drugs (SLID: AMK, KAN, CAP) and fluoroquinolones (FLQ: CIP, OFL, MOX) were also analysed. Insufficient phenotypic data was available for the inclusion of the new and repurposed drugs, bedaquiline, delamanid and linezolid. To reveal loci associated with drug resistance complementary methods were applied to mutations and aggregated non-synonymous mutations, a tree-based “PhyC” test for convergent evolution to detect homoplastic variants[10] and a GWAS approach within a mixed regression framework (See **Online methods**). Specifically, the low frequency of variants requires the aggregation of mutations to increase the power of detecting associated loci[15]**,** and a mixed model approach has been demonstrated to work well at adjusting for the confounding effects of *Mtb* lineage, sub-lineage and outbreak-based population structure[15]**.** A SNP-based GWAS was used to identify individual variants associated with drug resistance expected to fall within the genes found associated in the 'main' analysis. The phylogenetic-based "PhyC" test was applied to provide further evolutionary evidence. We report all findings that are below a calculated permutation threshold of $P<1x10^{-5}$. Some co-resistance associations were also revealed and annotated. Such findings may be expected to result from exposure to multiple anti-tuberculous drugs and the step-wise accumulation of mutations. Unless stated otherwise, all analysis used the complete dataset. First, we consider MDR-TB and XDR-TB phenotypes **(Table 1)** and then individual drug GWAS and evolutionary results (**Table 2**).

Gene and SNP-based GWAS and convergent evolution test for MDR-TB and XDR-TB

The gene-based GWAS of MDR-TB versus susceptible identified *rpoB* (RIF), *Rv1482c-fabG1* operon (INH, ETH), *inhA* (INH, ETH) and *katG* (INH). The *katG* mutations at codon 315 (S315T, S315N, S315R) were all statistically significant, and collectively were the most frequent mutations (81.7%) across all resistance loci identified, consistent with a recent study[16] and highlighting their pivotal role in the emergence of INH resistance and MDR-TB. The *katG* S315T mutation is thought to emerge before RIF resistance associated mutations and therefore, from an evolutionary standpoint, preclude the emergence of MDR-TB[16,17]. However, our analysis highlighted that *Rv1482c-fabG1* and *inhA* mutations, in the absence of *katG* S315T, can emerge prior to MDR-TB, as previously shown in two phylogenetically-independent clades in Lisbon[18,19]. The other frequent MDR-TB mutations in our study included *rpoB*-S450L (RIF, 67.0%), *embB*-M306L/V/I (EMB, 57.6%), and *rpsL*-K43R (STM, 45.8%) (**Supplementary table 3**), and the magnitude correlates with historical treatment practice and emergence of resistance. There are corresponding signals of INH/RIF co-resistance with other first-line drugs, with the detection of association signals for *rpsL* (STM), *embC-embA* intergenic region (EMB) and *embB* (EMB). SNP-based PhyC analysis detected the above loci, but in addition *folC* (PAS), *pncA-Rv2044c* intergenic region (PZA), and *whiB6-Rv3863* intergenic (putative STM or ETH) regions.

The gene-based GWAS of XDR-TB versus MDR-TB identified mutations in *gyrA* (FLQ), *rpoB* (RIF), *rrs* (aminoglycosides) and *ubiA* (EMB). One *ubiA* mutation (T180V, EMB) has not been previously reported and was found using the SNP-based GWAS approach. The PhyC test additionally revealed *eis-Rv2417c* (KAN), *gyrB (FLQ), rrs (*aminoglycosides) SNPs, and a novel mutation in the *thyX-hsdS.1* intergenic region (A-9T, PAS).[20,21]

The gene-based GWAS comparing XDR-TB to susceptible identified *rpoC* (a compensatory mechanism for RIF resistance), *oxyR'-ahpC* (compensatory mechanism for INH), *ethA* (ETH), *ethA-ethR* intergenic region (ETH), *eis-Rv2417c* (KAN) and *PPE52-nuoA* (novel intergenic region, G-314T). The PhyC test additionally detected SNPs in *gyrB* (FLQ, D461N, D641H, T500N, T500I and A504V)*,* supported the *thyX-hsdS.1* intergenic region SNP finding (PAS, A-9T), as well as endorsing the *ubiA* SNP associations (EMB, V188A, A249T). The *drrA* Arg262Gly mutation was significantly associated with XDR-TB compared to susceptible (mutation frequency 19% vs. 0%, respectively, $P<2x10^{-9}$). We hypothesize that *drrA* may be involved in export of drugs across the membrane based on its strong association with XDR-TB in our study and its functional annotation as a probable transporter of antibiotics across the membrane (http://tuberculist.epfl.ch). This hypothesis is in accordance with the findings that *rpoB* mutations in *Mtb* may trigger compensatory transcriptional changes in secondary metabolism genes, in particular, in the biosynthesis and export of phthiocerol dimycocerosate (PDIM), increasing its expression and activity. As a consequence these strains became more virulent and multidrug resistant, increasing their fitness by increased efflux activity and lipid metabolism[22,23].

Similarly, a mutation in the *Rv1144-mmpL13a* intergenic region (C-102A) was highly associated with XDR-TB versus susceptible (mutation frequency 19% vs. 0%, respectively, $P<7x10^{-8}$). This mutation sits in the promoter to the operon containing *mmpL13a* and *mmpL13b,* which code for transmembrane transport proteins and could influence expression of these proteins[7].

Lineage-specific and compensatory mechanisms

We conducted a stratified GWAS per lineage to identify lineage-specific loci associated with drug resistance. The majority of associations were present in more than one lineage. The largest number of lineage-specific drug resistance mutations were found in lineage 4, which was the largest collection investigated and contained more genetically diverse clones[6], implying that geographically restricted mutations are being captured (**Supplementary table 4**). A previously unreported putative compensatory locus was identified for pyrazinamide (*pncB1*) through analysis of lineage 1 which reached borderline significance for lineage 3.

We applied a systematic approach to reveal epistatic interactions between GWAS loci (from **Table 2**) or explore known compensatory effects using a test of non-random association to detect the frequent co-occurrence of mutations in pairs of loci (Fisher exact test, P-value cut-off $<10^{-8}$) (**Supplementary table 5**). Deep phylogenetic mutations were removed to increase robustness. This approach proved to be successful at identifying well-known compensatory relationships between *rpoB* and *rpoC* loci (RIF)[8], *rpoB* and *rpoA* (RIF)[24] and *katG* and *oxyR'-ahpC* (INH)[25]. We captured the frequent co-occurrence of *embB* and *ubiA* mutations which together are known to lead to high levels of EMB resistance[26], and they are therefore unlikely to represent a compensatory mechanism. Novel epistatic relationships included *pncA* with *pncB2* (PZA) and *thyA* with *thyX-hsdS.1* (PAS). The *pncB2* effect appears to be specific to lineage 4 **(Supplementary table 6)**. The other nicotinamide co-factor, *pncB1,* had weaker evidence of an epistatic relationship with *pncA* in lineage 1 (P=0.0016) **(Supplementary table 6).** Similarly, there was marginal evidence for *pyrG* (lineage 4, P=0.00016)[27] and *Rv0565c* (lineage 2, P=0.00027) with *ethA* (ETH)[28] **(Supplementary table 6)**. Follow-up investigations will need to determine whether mutations in these loci have an impact on the MIC or function as compensatory mechanisms.

Overall, the GWAS approach was effective at detecting known drug resistance determinants and epistatic (gene-gene) relationships and identified novel ones that warrant functional validation in future studies. As resistance loci for individual drugs, especially second-line treatments, may be masked by an analysis of the composite MDR-TB and XDR-TB outcomes, we repeated the GWAS, PhyC test and epistatic analysis for the 14 individual drugs considered.

Gene and SNP-based GWAS and convergent evolution test for individual drugs

Rifampicin, isoniazid and ethionamide

The *rpoB* locus showed the strongest association with RIF resistance, but the compensatory effects of *rpoC* and *rpoA* were also evident through homoplasy SNP analysis. As previously reported non-synonymous SNPs in *rpoC* (272 identified) were spread across the whole gene[29]. Altered or diminished activity of the catalase-peroxidase enzyme *KatG* is the most frequent mechanism of isoniazid resistance[30]**,** and as expected, the *katG* gene ranked first in the GWAS for this drug. Mutations in proposed INH drug targets, *kasA* and *kasB* previously included in some drug resistance databases, did not reach statistical significance in our study[31], suggesting an odds ratio below our detection level of 1.4 (with 99% confidence of detection, 90% statistical power)**.** Both *inhA,* encoding the molecular target of isoniazid[32] and the *Rv1482c-fabG1* intergenic region harbouring its promoter, showed strong associations with INH and ETH, with greater effects in the former. In addition, *oxyR'-ahpC* intergenic associated mutations (20 detected) were found in the presence of *katG* polymorphisms (28), supporting its role as a compensatory mechanism. For ethionamide, the *ethA* locus, encoding the drug-metabolising enzyme was found to be associated with resistance as described previously[33]. A total of 153 non-synonymous mutations were identified in *ethA,* scattered throughout the gene and mostly affecting codons different from those already described[9]**.**

Ethambutol

Mutations in the *embCAB* operon, which encodes for enzymes involved in the biosynthesis of arabinan components of the mycobacterial cell wall, are mostly responsible for EMB resistance but are not fully penetrant for resistance[34]**.** The *embB* and the *embC-embA* intergenic region had the strongest associations. *Rv3806c* (*ubiA*), described to contribute to high levels of EMB resistance *in vitro*[17] was also significantly associated in our analysis demonstrating a role in clinical samples too across all four lineages. Two novel loci were identified: *Rv2820c* thought to enhance mycobacterial virulence *ex vivo* and *in vivo,* and *Rv3300c* a conserved protein with unknown function (http://tuberculist.epfl.ch).

Pyrazinamide

The *pncA* locus was the highest ranked association with PZA resistance in the GWAS and was a target of independent mutation, consistent with its established role[35]**.** Additionally, many low frequency SNPs were reported across the whole gene which were not used in the association analysis and could potentially confer resistance (**Supplementary data 1**). Other proposed PZA targets, namely *rpsA*[36] and *panD*[37], did not reach statistical significance in the GWAS and were not targets of independent mutation among PZA resistant strains in our collection.

Streptomycin

The rpsL, *rrs* and *gid* loci, all known to be involved in STM resistance[20] were identified by GWAS. Mutations in *rpsL* are known to lead to high levels of STM resistance[38], and accordingly we observed high odds ratios indicative of high penetrance in association signals in this locus (**Figure 3**). In contrast, candidate *rrs* and *gid* gene polymorphisms showed weaker signals (lower odds ratio) in the overall GWAS, which concurs with existing evidence that *gid* and *rrs* mutations confer low levels of resistance[38]. When considering the odds ratios across all SNP-drug associations, those from *rrs* and *gid* were much lower on average than those from *pncA* (PZA) and *katG* (INH) (**Figure 3**). This analysis demonstrates a potential utility of using odds ratios and their statistical significance to indicate the impact of a mutation and its propensity to cause low, intermediate or high level resistance.

Fluoroquinolones and Second-line injectables

The gene- and SNP-based GWAS analysis revealed the *gyrA* locus, which encodes for the molecular target of FLQ[39], as the strongest association signal. In addition to homoplastic mutations in *gyrA*, evidence of independent mutation was detected in *gyrB*[40]. The *Rv2688c* C213R mutation was associated with MOX and FLQ resistance but did not reach statistical significance in OFL. The antibiotic transport ATP-binding protein encoded by *Rv2688c* is a known FLQ efflux gene[41]**.** As expected the strongest resistance gene and SNP-based association signals across AMK, KAN, and CAP was with the aminoglycoside (SLID) target gene *rrs*[20]**.** Association was observed with mutations in the *eis* promoter known to result in low levels of KAN resistance but not in co-resistance with other aminoglycosides[42]**.** The median odds ratio for *eis* promoter mutations is lower than that of *rrs* mutations (**Figure 3**), further supporting that *rrs* mutations confer higher levels of KAN resistance.

D-Cycloserine

CYS inhibits the Alr enzyme, responsible for the conversion of L-Alanine into D-Alanine, by competing with L-Alanine for the active site. Resistance to CYS results from mutations in the *alr* coding region[43]**.** In our study *alr* was significantly associated with CYS resistance (**Table 2**) in line with recent evidence showing that clinical strains with *alr* mutations exhibit increased resistance to CYS[12] and harboured multiple homoplastic mutations including Phe4Leu, Lys113Arg and Met343Thr. In a previous study, the Met343Thr mutation was detected in an XDR-TB strain that had been exposed to CYS treatment, predicted to alter the protein structure of Alr, and therefore it was hypothesised to be involved in CYS resistance[44]. To further understand the functional impact of the mutations found in *alr* we modelled the effect of these variants using the available crystal protein structure (PDB 1XFC, **Supplementary figure 3**). Mutations in *alr* were found to differ in their proximity to the CYS binding site and their effect on protein stability and ligand binding (**Supplementary table 7**). The Met343Thr mutation (found in 12 susceptible and 2 resistant isolates) was predicted to have more drastic effect on protein structure compared to Lys113Arg, the most frequent mutation among CYS resistant isolates (in 7 susceptible and 23 resistant isolates). There appears to be a balance between the fitness cost associated with mutations and their frequency (**Supplementary table 7)**. The Met343Thr mutation appears independently throughout the phylogenetic tree, but did not reach statistical significance for association to drug resistance (XDR-TB or CYS), implying that selection may be acting on this mutation but drug resistance may not be the driving factor.

Para-aminosalicylic acid

PAS is a pro-drug that is converted into its active form by *thyA* - a thymidylate synthase, which is an essential gene for *Mtb* survival. The candidate drug resistance loci are those involved in folate metabolism and biosynthesis of thymidine nucleotides (*thyA*, *dfrA*, *folC*, *folP1*, *folP2* and *thyX*)[21]**.** Of these, *thyA* and *thyX-hsdS.1* (directly upstream of *thyX*) and were found to be associated with PAS drug resistance in both gene- and SNP-based GWAS. Importantly, it has been shown that G-16A SNP found in our study increased *thyX* expression by 18-fold relative to wild-type promoter although no link with PAS resistance was made[20]**.** Of 3 PAS resistance strains with the G-16A *thyX* promoter mutation, 2 also had a *thyA* mutation (P145L, H207R), further supporting that up-regulation of *thyX* is involved in resistance to PAS[28], or has a compensatory role. The G-16A *thyX* is a homoplastic mutation, and therefore more likely to be compensatory.

The odds ratios for the novel findings were less than those for known ones (present in the TBProfiler database), reflecting that the ability of the GWAS to discover effect sizes of lower magnitude **(Figure 3)**. However, novel SNPs associations for PZA and RIF were more likely to have higher odds ratios. A pathway analysis comparing MDR-TB/XDR-TB to susceptible strains revealed only one significant annotation cluster with 17.7-fold enrichment for antibiotic resistance and response to antibiotics ($P<2x10^{-7}$), further confirming the robustness of the GWAS approach.

Association analysis using small indels and large deletions

An analysis of genome-wide small indels revealed associations in candidate resistance genes and operons (**Supplementary table 8, Supplementary data 1**). The candidate genes differed in their abundance of small indels, reflecting their essentiality for survival: drug targets had less density of indels whereas drug-metabolising enzymes had a greater density. For example, the *pncA* gene was the most polymorphic coding region (PZA, 44.72 indels /kb) while the least polymorphic was *rpoB* (RIF, 2.3 indels /kb). Although, the majority of small indels (83%) in the candidate regions were 1bp in length and caused frame-shifts, the indels in *rpoB* inserted or deleted whole codons, i.e. they did not cause a shift in the codon reading frame. Indels in *rpoB, pncA* and the *embAB* promoter region were associated with MDR-TB, XDR-TB and their respective targets/activators. Indels in *ethA* were associated with ETH and XDR-TB resistance. Similarly, *gid* indels were associated with STM as expected.

The analysis of CYS revealed indel associations with the *ald* gene, supporting recent reports that loss of function in *ald* confers resistance[12]. Thus resistance to CYS appears to be conferred by both SNPs in *alr* and indels in *ald*. Indels found in *rrs* were associated with KAN and CAP resistance, however they did not reach statistical significance for STM, which has a different drug binding site. CAP resistance was also found to be associated with three indels in *tlyA,* two of which are located at the 3' end of the gene. In general, indels were distributed throughout the gene lengths however there was some evidence of areas of higher density such as the *pncA* region between codons 130 and 132 (close to the catalytic centre) and the *rpoB* 427-434 codon region.

The only large deletion association identified by GWAS was a region encompassing the *thyA* and *dfrA* genes and PAS resistance. Five samples across 4 countries contained large *thyA-dfrA* deletions of varying length **(Supplementary table 9, Supplementary figure 4).** Associations in partial or whole gene deletions in *katG, ethA* and *pncA,* were close to statistical significance (P<0.05). These genes activate pro-drugs, and none are considered to be essential to *Mtb* survival. The large deletions detected occur independently in different branches of the phylogenetic tree and are likely to offer an alternative route to resistance compared to small genomic variants, across lineages and populations.

Effects on predicting drug resistance phenotypes with new SNPS and indels from GWAS

We sought to establish if any of the mutations found in association and homoplastic analysis increased the predictability of individual drug resistance phenotypes (**Table 3**). We used the reported phenotypic drug susceptibility test result as the reference standard to calculate the sensitivity and specificity for mutation-resistance predictions. Using a previously established library of mutations[9,19] (TBDR library), we found that although the sensitivity was greater than 80% in 8/14 drugs, a substantial proportion of resistance phenotypes were not explained by known mutations, particularly in second-line drugs. Using the novel SNPs identified in this study we gained sensitivity for STM (+1%), PAS (+10%), and CYS (+50%, not included in the TBDR library) (**Table 3**). The additional inclusion of small indels and large deletions further improved the predictive ability for 9 drugs while maintaining specificities of >90%, except for ETH which is 70% (**Table 3**).

DISCUSSION

To provide genomic insight into *Mtb* drug resistance we have combined the power of whole genome sequencing with a genome-wide association analytical approach in the largest and most geographically widespread study to date, encompassing a total of 6,465 clinical isolates of *Mtb* from more than 30 countries. Large sample sizes are required to identify complex or infrequent genetic effects, but also to negate effects due to possible errors in phenotypic drug susceptibility testing and misclassification[45]**.** The lack of standardization of phenotypic testing methodologies for *Mtb* is also a potential source of bias which was reduced by the inclusion of samples from a number of different countries and laboratories using a variety of quality assured testing methodologies. A recent large study demonstrated the usefulness of the convergent approach to detect known mutations underlying MDR-TB[16]. However, the incompleteness of first- and second-line drug resistance outcomes meant association analysis to detect novel mutations was not possible. Whilst resistant phenotypes may be imputed from established resistance causing mutations, inferring susceptibility to a drug cannot be assumed in the absence of corroborating evidence[19]. The completeness of our susceptibility test data meant that both GWAS and homoplasy-based methods could be applied across 14 drugs. The predominance of *Mtb* genome polymorphisms of low frequency required the adoption of a robust rare-variant approach, where mutations were aggregated by gene and operon (a surrogate for pathways)[15]. However, the large sample size enabled us to detect mutations at low frequency. The GWAS identified well-established resistance loci and compensatory relationships, thereby confirming the authenticity and robustness of the approach. It also revealed several recently discovered loci (*folC*, *ubiA*, *thyX-hsdS.1*, *thyA*, *alr*, *ald*, *dfrA-thyA*), new epistatic relationships (*pncA* with *pncB2*, and *thyA* with *thyX-hsdS.1*) and efflux pumps represented by the ABC transporters *drrA* and *Rv2688c* associated with drug resistance. The novel genetic markers associated with resistance identified in this GWAS included SNPs in the *ethA* and *thyX* promoters, small indels in *pncA* and *ald,* and large deletions in pro-drug activators such as *ethA* and *katG*. These loci warrant functional follow-up and characterization studies to fully elucidate their role in treatment failure. The associations identified may shed light on the molecular mechanisms underlying drug resistance and assist in the design of novel antibiotics.

Improved knowledge of the molecular mechanisms responsible for resistance to drugs used to treat MDR-TB and XDR-TB is important. Second-line drug susceptibility testing is technically challenging to perform and quality assured testing is not widely available in all TB endemic countries. Such countries also tend to have deficient directly observed treatment, short-course (DOTS) programs and consequently are at risk of high rates of resistance. In our study, sample sizes for second-line drugs were reduced compared to the first-line drugs. This was due to the lower prevalence of resistance to second-line drugs and the fact that isolates susceptible to first-line drugs are not routinely tested for second-line drugs.

However, due to the large effect that causal mutations have on drug resistance phenotypes, although not ideal, relatively small samples of bacterial genomes can be sufficient to identify causal mutations [43] as has been demonstrated in previous studies on *Mtb* [10-12]. It should be noted that bedaquiline, delamanid and linezolid were excluded from our analysis due to the paucity of phenotypic susceptibility data.

The analysis also highlighted the importance of indels on drug resistance, particularly their high density in drug-metabolizing genes, in contrast to highly essential drug-target genes where their density was low. The inclusion of small indels and large deletions improved the predictability of resistance phenotypes. However, for drugs like CYS and PAS mechanisms of drug resistance remain unknown and larger numbers of resistant cases will be required to elucidate them. It is also possible that unknown mechanisms may be explained by the role of epigenetics and gene expression[46].

Mtb strains are usually classified as drug resistant or susceptible based on their capacity to grow *in vitro* when exposed to a critical concentration of the drug. Phenotypic testing methods have a degree of uncertainty, especially close to the threshold[45]**.** Testing against a range of drug concentrations to establish the minimum inhibitory concentration (MIC) is a preferred approach but is not routinely undertaken[42]. Most resistance is of a high level when strains can survive high drug concentrations but intermediate and low levels of resistance are also reported for some drugs, and in such cases, increased dosing may be beneficial for the patient[47]**.** MIC values were not available for every isolate presented here, but despite this limitation, loci known to be involved in low-levels of resistance (**Table 3**), were identified by our analysis. Indeed, our analysis revealed a relationship between known levels of resistance and the odds ratios from the GWAS, which could aid the clinical interpretation of molecular diagnostic data including measuring the sensitivity and specificity of individual mutations when diagnosing drug resistance.

Emergence of resistance is driven by drug exposure and local TB treatment practices are a major influence on the prevalence and pattern of resistance. A limitation of this study was the sampling methodology since collection of the isolates was not controlled or systematic and resistant isolates were not evenly distributed across collection sites. However, within our study population we covered the four major *Mtb* lineages across 5 continents and sampled multiple geographical regions, allowing us to observe differences in the prevalence of drug resistance mutations and mechanisms. Some of drug resistance and compensatory/epistatic relationships were found to vary across geographical populations and bacterial lineage, implying that regional variation should be considered to fully characterise genotype-phenotype relationships. The differential lineage effects could impact on relative virulence between strain-types. Enhanced understanding of the genetic basis of anti-tuberculous phenotypic drug resistance will also aid in the development of more accurate molecular diagnostics for drug-resistant TB. An important finding of this study is the significance of genomic variation other than SNPs which has implications for the design of molecular tests for resistance. Improved tools are needed to guide treatment of patients with multidrug-resistant disease where personalised treatment offers improved rates of cure[48]**.** Next generation sequencing offers a comprehensive assessment and may be used to guide treatment[48]. Although such technology is currently being implemented in some low burden countries such as the United Kingdom, it remains to be trialled in resource-poor settings that are representative of the majority of TB patients worldwide.

Section 1.01 ONLINE METHODS

(a) Sequence data and variant calling

Sequence data for 6,465 *M. tuberculosis complex* clinical isolates were generated as part of a collaborative global drug resistance project (n=2,637, pathogenseq.lshtm.ac.uk) or downloaded from the public domain (n=3,828) (**Supplementary table 1**). All isolates had undergone drug susceptibility testing by phenotypic methods. These isolates represented multiple populations from different geographic areas, and all four main lineages (1 to 4) (**Supplementary table 1**). The 2,637 samples not previously sequenced were Illumina sequenced generating paired-end reads of at least 50 bp with at least 50-fold genome coverage. The analytical workflow for the raw sequence data is summarised in **Supplementary figure 5**. The new and archived raw sequence data were aligned to the H37Rv reference genome (Genbank accession number: NC_000962.3) using the *BWA mem* algorithm[49]**.** The *SAMtools/BCFtools*[50] and *GATK*[51] software was used to call SNPs and small indels using default options. The GATK parameters used are "-T UnifiedGenotyper -ploidy 1 -glm BOTH -allowPotentiallyMisencodedQuals 2". The overlapping set of variants from the two algorithms was retained for further analysis. Alleles were additionally called across the whole genome (including SNP sites) using a coverage-based approach[6,15]**.** A missing call was assigned if the total depth of coverage at a site did not reach a minimum of 20 reads or none of the four nucleotides accounted for at least 75% of the total coverage. Samples or SNP sites having an excess of 10% missing genotype calls were removed. This quality control step was implemented to remove samples with bad quality genotype calls due to poor depth of coverage or mixed infections. The final dataset included 6,465 isolates and 102,160 genome-wide SNPs. *Delly2* software[52] was used to find large deletions. All large deletions were confirmed using localised *de novo* assembly, and those found in association analysis (*dfrA/thyA, pncA, ethA/ethR, katG*) confirmed using PCR.

Phenotypic drug susceptibility testing

Drug susceptibility data was obtained from World Health Organisation recognised testing protocols[53]**.** The *M. tuberculosis* isolates that provided sequence data included in this study are summarised in **Supplementary table 1.** Each sequence included in the study was derived from an isolate from an individual patient. Some DNA samples were from archived stocks (e.g. India, collected prior to 2009 and Malawi, collected between 1996 and 2010) and others were extracted specifically for this study. Information regarding isolates with previously reported sequence data was derived from published materials. Isolates were classed as resistant or susceptible to a drug on the basis of phenotypic testing using either the BACTEC 460 TB System (Becton Dickinson), the BACTEC Mycobacterial Growth Indicator Tube (MGIT) 960 system (Becton Dickinson)[54], solid agar or Lowenstein Jensen slopes[55,56]**.** Not all samples were tested for resistance to all drugs, most notably some isolates found susceptible to the first-line drugs were not subjected to testing for resistance to second-line drugs. Where isolates were not tested for resistance to a particular drug they were excluded from the analysis for that drug. Drug susceptibility testing was mainly undertaken in local laboratories participating in the WHO supranational laboratory network using the recognised testing protocols[53]**.** Isolates from Malawi were shipped to the United Kingdom's Mycobacterium Reference Laboratory for testing. Isolates from Uganda were tested at the Joint Clinical Research Centre (JCRC) in Kampala with quality control performed by the US Centers for Disease Control and Prevention (CDC). The Peruvian isolates were initially tested for resistance to rifampicin and isoniazid using the Microscopic Observation Drug Susceptibility assay (MODS)[56] at the Universidad Peruana Cayetano Heredia (UPCH) prior to transfer to the national reference laboratory for further testing. In Peru susceptibility to pyrazinamide (PZA) was assessed by the Wayne assay; a colorimetric biochemical test during which PZA is hydrolysed to free pyrazinoic acid[57]. Testing using the BACTEC 960® MGIT® or BACTEC 460® (Becton-Dickinson®) was performed according to the manufacturer's indications[58]**.** Pyrazinamide sensitivity was determined by using BACTEC 7H12 liquid medium, pH 6.0, at 100 µg/mL (BACTEC PZA test medium, Becton Dickinson). When testing on agar critical drug concentrations used were rifampicin 1 µg/mL, isoniazid 0.2 µg/mL, streptomycin 2 µg/mL, and ethambutol 5 µg/mL, ciprofloxacin 2 µg/mL, amikacin 5µg/mL, capreomycin 10 µg/mL, kanamycin 5 µg/mL (Pakistan 6 µg/mL), ethionamide 5 µg/mL and para-aminosalicylic acid 2 µg/mL[55]**.** For Lowenstein-Jensen drug concentrations used were for streptomycin 4.0 µg/ml, isoniazid 0.2 µg/ml, rifampicin 40.0 µg/ml, ethambutol 2.0 µg/ml, capreomycin 40.0 µg/ml, kanamycin 30.0 µg/ml (China) or 20.0 µg/ml (Vietnam), ofloxacin 2.0 µg/ml, ethionamide 40 µg/ml, thioacetone (10 µg/ml), pyrazinamide 200 µg/ml, cycloserine 30 µg/ml and *para*-aminosalicylic acid (PAS) 0.5 µg/ml[57].

(b) Phylogenetic tree and association analysis

The best-scoring maximum likelihood phylogenetic tree rooted on *Mycobacterium canettii* was constructed by *RAxML* software[59] (10,000 bootstrap samples) using the 102,160 high quality SNP sites. Spoligotypes were inferred *in silico* using *SpolPred*[60]**,** and strain-types determined using lineage-specific SNPs[6]**.** Further population structure assessment was performed using principal components analysis (**Supplementary figure 3)**, which clustered samples by genotype congruent with the phylogenetic tree. The principal components were calculated from a SNP pair-wise distance matrix between each sample, and the first five components (summarising 82.7% of genetic variation) were used as covariates in the regression-based association models. Mixed regression models were employed to estimate the strength of association between the binary drug resistance outcome (resistance vs. susceptible) and the aggregate number of mutations (SNPs, indels or large deletions) by coding region, RNA loci and intergenic regions, as well as operons[15]**.** The operons or functional units containing clusters of genes under the control of the same promoter were determined from TBDB (http://www.tbdb.org). Gene function was extracted from the Tuberculist webserver (http://tuberculist.epfl.ch**).** The mixed models also included the principal components to account for the main *Mtb* lineage and sub-lineage effects, and a SNP inferred kinship matrix as a random effect to account for highly related samples and fine-scale population structure due to potential outbreaks[15]**,** and were implemented in GEMMA (v.1.1.2) software[61]**.** To minimise any co-resistance between drugs, and we adjusted for the presence of other resistance in the regression models. Statistical significance thresholds to account for multiple testing were established using a permutation approach that sorted phenotypic test data without replacement and re-performed GWAS analysis (10,000 times). The determined P-value threshold was $1x10^{-5}$. All statistical analysis was performed using R software. To identify SNPs enriched by convergent evolution, the *phyC* approach was employed[10] using the implementation made available in a previous study[62]. Any potential co-resistance effects were dissected through consulting gene annotation and published literature to report the most plausible role in drug resistance. Additionally, long branches in the phylogenetic tree leading up to clades enriched with drug resistant isolates leads to spurious associations. Truly drug resistant mutations often originate multiple times independently in the phylogeny. Mutations which originated once in the tree (i.e. clade-specific mutations), which are likely to lead to spurious associations, were removed from the GWAS results.

Detection of putative compensatory mechanisms

Loci were identified as putative compensatory loci if they: (i) were associated with drug resistance, (ii) harboured homoplastic mutations, (iii) shared a similar biological function with a known drug-target or drug-activating enzyme, and (iv) were significantly more mutated in the presence of mutations in the drug-target or drug-activating enzyme coding gene. In the latter, deep phylogenetic and synonymous SNPs were removed prior to calculating the number of samples with non-synonymous SNPs at genes of interest (e.g. Ala1075Ala at *rpoB* or Glu1092Asp at *rpoC*). The significance of differences between studied genes was calculated using Fisher's exact test ($P<10^{-8}$).

Protein mutation modelling

Apo crystal structures for *alr* were downloaded from the Protein Data Bank (PDBe1XFC[63]) and then subjected to modelling of missing residues, WinCOOT regularisation, and removal of pyridoxal 5′-phosphate from both chains. The mCSM (http://structure.bioc.cam.ac.uk/mcsm) and DUET (http://structure .bioc.cam.ac.uk/duet) web servers were used to assess changes in protein stability, mCSM-PPI (http://structure.bioc.cam.ac.uk/mcsm_ppi) to quantify effects on protein-protein interactions and mCSM-Lig (http://bleoberis.bioc.cam.ac.uk/mcsm_lig) to quantify effects on drug binding[64–66]**.** For ligand binding, D-Cycloserine was docked in the active site using Autodock Vina and Gold software[67,68].

DATA AVAILABILITY

All raw sequencing data are available, and the study accession numbers are listed in **Supplementary table 1.** The phenotypic data are available from the study website (http://pathogenseq.lshtm.ac.uk/#tuberculosis).

ACKNOWLEDGMENTS

The project was supported by the KAUST faculty baseline research fund (KAUST-BRF-BAS/1/1020-01-01) to AP; The authors wish to thank members of KAUST Bioscience Core laboratory who sequenced the samples. We thank the Wellcome Trust Sanger Institute core and pathogen sequencing and informatics teams who were involved in the Malawi and Uganda studies. The work was funded in part by the Wellcome Trust (Grant numbers

WT096249/Z/11/B, WT088559MA, WT081814/Z/06/Z, and WT098051), and the Wellcome Trust-Burroughs Wellcome Fund Infectious Diseases Initiative grant (number 063410/ABC/00/Z). FC was the recipient of a Bloomsbury College PhD Studentship and was supported by the Wellcome Trust (201344/Z/16/Z); JPe received a *Fundação* para a *Ciência* e a *Tecnologia* (Portugal) Post-doctoral fellowship fund (SFRH/BPD/95406/2013). The Calouste Gulbenkian Foundation, the Institute Gulbenkian in Lisbon and European Society of Clinical Microbiology and Infectious Diseases supported the research of CP, JPe, IP and MV. JPh is funded by a BBSRC PhD studentship. TGC is funded by the Medical Research Council UK (Grant no. MR/K000551/1 and MR/M01360X/1, MR/N010469/1). TM is supported by the Ministry of Health, Labor and Welfare of Japan (H21- Shinkou-Ippan-008 and H24-Shinkou-Ippan-010). We thank Nerges Mistry (Foundation for medical research, Mumbai) for contributing *M. tuberculosis* archived strains and drug-sensitivity testing data. We wish to thank Prof. Goncalo Moniz at the Laboratorio Central de Saude Publica for supporting the collection of samples in Brazil, and the South African National Health Laboratory Service for their contribution providing access clinical *Mtb* isolates. The MRC eMedLab computing resource was used for bioinformatics and statistical analysis. The authors declare no conflicts of interest. The work has been performed as part of the TB Global Drug Resistance Collaboration (http://pathogenseq.lshtm.ac.uk/#tuberculosis).

AUTHOR CONTRIBUTIONS

RM, AP and TC conceived and directed the project. KM coordinated sample collection and undertook DNA extraction. SAl, AOA, AA, TB, MC, Ach, AC, KD, LG, JG, DH, RH, ZH, PH, MJ, EJ, TM, AM, NM, DM, SP, IP, CP, JPe, JR, PS, NS, CS, EOS, ES, NP, MV and RW undertook sample collection, DNA extraction, genotyping and phenotypic drug resistance testing. MBN, MAS, ZR and SA prepared libraries for Illumina sequencing. JPa led the generation of Malawian and Ugandan sequencing data. FC, JPh and GH performed bioinformatic and statistical analyses under the supervision of AP and TC. SP and YO performed additional confirmatory analysis under the supervision of MH, NF and TC. FC, JPh, SP, NF, MH, RM, AP, and TC interpreted results. FC, JPh, RM and TC wrote the first draft of the manuscript. TB, GH, MC, JG, PH, EJ, JPa, JPe, MH, NF, SP, JR, CS, ES, MV, RW, and AP commented and edited on various versions of the draft manuscript and all authors approved the manuscript. FC, JPh, RM, AP and TC compiled the final manuscript.

DISCLOSURE DECLARATION

There are no conflicts of interest.

REFERENCES

1. Dheda, K. *et al.* Global control of tuberculosis: from extensively drug-resistant to untreatable tuberculosis. *Lancet. Respir. Med.* **2,** 321–38 (2014).
2. Dheda, K. *et al.* Outcomes, infectiousness, and transmission dynamics of patients with extensively drug-resistant tuberculosis and home-discharged patients with programmatically incurable tuberculosis: a prospective cohort study. *Lancet Respir. Med.* (2017). doi:10.1016/S2213-2600(16)30433-7
3. Bastos, M. L. *et al.* Treatment outcomes of patients with multidrug-resistant and extensively drug-resistant tuberculosis according to drug susceptibility testing to first- and second-line drugs: an individual patient data meta-analysis. *Clin. Infect. Dis.* **59,** 1364–74 (2014).
4. Shean, K. *et al.* Drug-Associated Adverse Events and Their Relationship with Outcomes in Patients Receiving Treatment for Extensively Drug-Resistant Tuberculosis in South Africa. *PLoS One* **8,** e63057 (2013).
5. Clark, T. G. *et al.* Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing. *PLoS One* **8,** e83012 (2013).
6. Coll, F. *et al.* A robust SNP barcode for typing Mycobacterium tuberculosis

complex strains. *Nat. Commun.* **5,** 4812 (2014).

7. Black, P. A. *et al.* Energy metabolism and drug efflux in Mycobacterium tuberculosis. *Antimicrob. Agents Chemother.* **58,** 2491–503 (2014).
8. de Vos, M. *et al.* Putative compensatory mutations in the rpoC gene of rifampin-resistant Mycobacterium tuberculosis are associated with ongoing transmission. *Antimicrob. Agents Chemother.* **57,** 827–32 (2013).
9. Coll, F. *et al.* Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med.* **5:51,** (2015).
10. Farhat, M. R. *et al.* Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. *Nat. Genet.* **45,** 1183–9 (2013).
11. Zhang, H. *et al.* Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance. *Nat. Genet.* **45,** 1255–1260 (2013).
12. Desjardins, C. A. *et al.* Genomic and functional analyses of Mycobacterium tuberculosis strains implicate ald in D-cycloserine resistance. *Nat. Genet.* **48,** 544–551 (2016).
13. Earle, S. G. *et al.* Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat. Microbiol.* **1,** 16041 (2016).
14. Chewapreecha, C. *et al.* Comprehensive Identification of Single Nucleotide Polymorphisms Associated with Beta-lactam Resistance within Pneumococcal Mosaic Genes. *PLoS Genet.* **10,** e1004547 (2014).
15. Phelan, J. *et al.* Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC Med.* **14,** 31 (2016).
16. Manson, A. L. *et al.* Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. *Nat. Genet.* (2017). doi:10.1038/ng.3767
17. Cohen, K. A. *et al.* Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of Mycobacterium tuberculosis Isolates from KwaZulu-Natal. *PLOS Med.* **12,** e1001880 (2015).
18. Perdigão, J. *et al.* Unraveling Mycobacterium tuberculosis genomic diversity and evolution in Lisbon, Portugal, a highly drug resistant setting. *BMC Genomics* **15,** 991 (2014).
19. Phelan, J. *et al.* The variability and reproducibility of whole genome sequencing technology for detecting resistance to anti-tuberculous drugs. *Genome Med.* **8,** 132 (2016).
20. Meier, A., Sander, P., Schaper, K. J., Scholz, M. & Böttger, E. C. Correlation of molecular resistance mechanisms and phenotypic resistance levels in streptomycin-resistant Mycobacterium tuberculosis. *Antimicrob. Agents Chemother.* **40,** 2452–4 (1996).
21. Zhang, X. *et al.* Genetic Determinants Involved in *p* -Aminosalicylic Acid Resistance in Clinical Isolates from Tuberculosis Patients in Northern China from 2006 to 2012. *Antimicrob. Agents Chemother.* **59,** 1320–1324 (2015).
22. Bisson, G. P. *et al.* Upregulation of the phthiocerol dimycocerosate biosynthetic pathway by rifampin-resistant, rpoB mutant Mycobacterium tuberculosis. *J.*

Bacteriol. **194,** 6441–52 (2012).

23. Chatterjee, A., Saranath, D., Bhatter, P., Mistry, N. & Thomson, A. Global Transcriptional Profiling of Longitudinal Clinical Isolates of Mycobacterium tuberculosis Exhibiting Rapid Accumulation of Drug Resistance. *PLoS One* **8,** e54717 (2013).
24. Comas, I. *et al.* Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nat. Genet.* **44,** 106–10 (2012).
25. Sherman, D. R. *et al.* Compensatory ahpC gene expression in isoniazid-resistant Mycobacterium tuberculosis. *Science* **272,** 1641–3 (1996).
26. Safi, H. *et al.* Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-β-D-arabinose biosynthetic and utilization pathway genes. *Nat. Genet.* **45,** 1190–7 (2013).
27. Mori, G. *et al.* Thiophenecarboxamide Derivatives Activated by EthA Kill Mycobacterium tuberculosis by Inhibiting the CTP Synthetase PyrG. *Chem. Biol.* **22,** 917–927 (2015).
28. Merker, M. *et al.* Whole genome sequencing reveals complex evolution patterns of multidrug-resistant Mycobacterium tuberculosis Beijing strains in patients. *PLoS One* **8,** e82551 (2013).
29. Casali, N. *et al.* Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat. Genet.* **46,** 279–286 (2014).
30. Zhang, Y., Heym, B., Allen, B., Young, D. & Cole, S. The catalase—peroxidase gene and isoniazid resistance of Mycobacterium tuberculosis. *Nature* **358,** 591–593 (1992).
31. Larsen, M. H. *et al.* Overexpression of inhA, but not kasA, confers resistance to isoniazid and ethionamide in Mycobacterium smegmatis, M. bovis BCG and M. tuberculosis. *Mol. Microbiol.* **46,** 453–66 (2002).
32. Banerjee, A. *et al.* inhA, a gene encoding a target for isoniazid and ethionamide in Mycobacterium tuberculosis. *Science* **263,** 227–30 (1994).
33. DeBarber, A. E., Mdluli, K., Bosman, M., Bekker, L. G. & Barry, C. E. Ethionamide activation and sensitivity in multidrug-resistant Mycobacterium tuberculosis. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 9677–82 (2000).
34. Telenti, A. *et al.* The emb operon, a gene cluster of Mycobacterium tuberculosis involved in resistance to ethambutol. *Nat. Med.* **3,** 567–570 (1997).
35. Scorpio, A. & Zhang, Y. Mutations in pncA, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus. *Nat. Med.* **2,** 662–667 (1996).
36. Shi, W. *et al.* Pyrazinamide inhibits trans-translation in Mycobacterium tuberculosis. *Science* **333,** 1630–2 (2011).
37. Shi, W. *et al.* Aspartate decarboxylase (PanD) as a new target of pyrazinamide in Mycobacterium tuberculosis. *Emerg. Microbes Infect.* **3,** e58 (2014).
38. Perdigão, J. *et al.* GidB mutation as a phylogenetic marker for Q1 cluster Mycobacterium tuberculosis isolates and intermediate-level streptomycin resistance determinant in Lisbon, Portugal. *Clin. Microbiol. Infect.* **20,** O278–O284 (2014).
39. Takiff, H. E. *et al.* Cloning and nucleotide sequence of Mycobacterium tuberculosis gyrA and gyrB genes and detection of quinolone resistance

mutations. *Antimicrob. Agents Chemother.* **38,** 773–80 (1994).

40. Kocagöz, T. *et al.* Gyrase mutations in laboratory-selected, fluoroquinolone-resistant mutants of Mycobacterium tuberculosis H37Ra. *Antimicrob. Agents Chemother.* **40,** 1768–74 (1996).

41. Pasca, M. R. *et al.* Rv2686c-Rv2687c-Rv2688c, an ABC Fluoroquinolone Efflux Pump in Mycobacterium tuberculosis. *Antimicrob. Agents Chemother.* **48,** 3175–3178 (2004).

42. Zaunbrecher, M. A., Sikes, R. D., Metchock, B., Shinnick, T. M. & Posey, J. E. Overexpression of the chromosomally encoded aminoglycoside acetyltransferase eis confers kanamycin resistance in Mycobacterium tuberculosis. *Proc. Natl. Acad. Sci.* **106,** 20004–20009 (2009).

43. Awasthy, D., Bharath, S., Subbulakshmi, V. & Sharma, U. Alanine racemase mutants of Mycobacterium tuberculosis require D-alanine for growth and are defective for survival in macrophages and mice. *Microbiology* **158,** 319–327 (2012).

44. Köser, C. U. *et al.* Whole-Genome Sequencing for Rapid Susceptibility Testing of *M. tuberculosis*. *N. Engl. J. Med.* **369,** 290–292 (2013).

45. Schön, T. *et al.* Mycobacterium tuberculosis drug-resistance testing: challenges, recent developments and perspectives. *Clin. Microbiol. Infect.* (2016). doi:10.1016/j.cmi.2016.10.022

46. Smith, T., Wolff, K. A. & Nguyen, L. Molecular biology of drug resistance in Mycobacterium tuberculosis. *Curr. Top. Microbiol. Immunol.* **374,** 53–80 (2013).

47. Cambau, E. *et al.* Revisiting susceptibility testing in MDR-TB by a standardized quantitative phenotypic assessment in a European multicentre study. *J. Antimicrob. Chemother.* **70,** 686–96 (2015).

48. McNerney, R. *et al.* Removing the bottleneck in whole genome sequencing of Mycobacterium tuberculosis for rapid drug resistance analysis: a call to action. *Int. J. Infect. Dis.* (2016). doi:10.1016/j.ijid.2016.11.422

49. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30,** 2843–2851 (2014).

50. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–9 (2009).

51. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).

52. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28,** i333–i339 (2012).

53. World Health Organization. WHO | Guidelines for surveillance of drug resistance in tuberculosis. (2009).

54. Kent, P. T. *Public health mycobacteriology: a guide for the level III laboratory*. (U.S. Dept. of Health and Human Services, Public Health Service, Centers for Disease Control, 1985).

55. CANETTI, G. *et al.* MYCOBACTERIA: LABORATORY METHODS FOR TESTING DRUG SENSITIVITY AND RESISTANCE. *Bull. World Health Organ.* **29,** 565–78 (1963).

56. Minion, J., Leung, E., Menzies, D. & Pai, M. Microscopic-observation drug susceptibility and thin layer agar assays for the detection of drug resistant tuberculosis: a systematic review and meta-analysis. *Lancet Infect. Dis.* **10,** 688–698 (2010).

57. Wayne, L. G. Simple pyrazinamidase and urease tests for routine identification of mycobacteria. *Am. Rev. Respir. Dis.* **109,** 147–51 (1974).

58. Palicova, F., Jahn, E. I. M. & Pfyffer, G. E. Susceptibility Testing of Mycobacterium tuberculosis to Anti-Tuberculosis Drugs: BACTEC ^{TM} MGIT ^{TM} 960 vs BACTEC ^{TM} 460TB System.

59. Stamatakis, A., Hoover, P. & Rougemont, J. A Rapid Bootstrap Algorithm for the RAxML Web Servers. *Syst. Biol.* **57,** 758–771 (2008).

60. Coll, F. *et al.* SpolPred: rapid and accurate prediction of Mycobacterium tuberculosis spoligotypes from short genomic sequences. *Bioinformatics* **28,** 2991–3 (2012).

61. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44,** 821–824 (2012).

62. Alam, M. T. *et al.* Dissecting vancomycin-intermediate resistance in staphylococcus aureus using genome-wide association. *Genome Biol. Evol.* **6,** 1174–85 (2014).

63. Velankar, S. *et al.* PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Res.* (2015). doi:10.1093/nar/gkv1047

64. Pires, D. E. V, Ascher, D. B. & Blundell, T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* **30,** 335–42 (2014).

65. Pires, D. E. V., Blundell, T. L. & Ascher, D. B. mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Sci. Rep.* **6,** 29575 (2016).

66. Pires, D. E. V, Ascher, D. B. & Blundell, T. L. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res.* **42,** W314-9 (2014).

67. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31,** 455–61 (2010).

68. Verdonk, M. L. *et al.* Modeling Water Molecules in Protein−Ligand Docking Using GOLD. *J. Med. Chem.* **48,** 6504–6515 (2005).

69. Wong, S. Y. *et al.* Mutations in gidB confer low-level streptomycin resistance in Mycobacterium tuberculosis. *Antimicrob. Agents Chemother.* **55,** 2515–22 (2011).

70. Rueda, J. *et al.* Genotypic Analysis of Genes Associated with Independent Resistance and Cross-Resistance to Isoniazid and Ethionamide in Mycobacterium tuberculosis Clinical Isolates. *Antimicrob. Agents Chemother.* **59,** 7805–7810 (2015).

71. Kambli, P. *et al.* Correlating rrs and eis promoter mutations in clinical isolates of Mycobacterium tuberculosis with phenotypic susceptibility levels to the second-line injectables. *Int. J. Mycobacteriology* **5,** 1–6 (2016).

Figure 1. Geographical distribution of the 6,465 *M. tuberculosis* isolates analysed in the study

This world map shows the main geographical origins of the *M. tuberculosis* isolates included in this study. The study comprises strains from more than 30 countries, of which the 18 major contributors are showed in this map. See **Supplementary table 1** for a detailed description of each dataset. Inner pie charts show the proportion of each of the main four lineages, and the outer charts the drug resistance phenotypes. 'Drug-resistant' refers to non-MDR-TB/XDR-TB resistance.

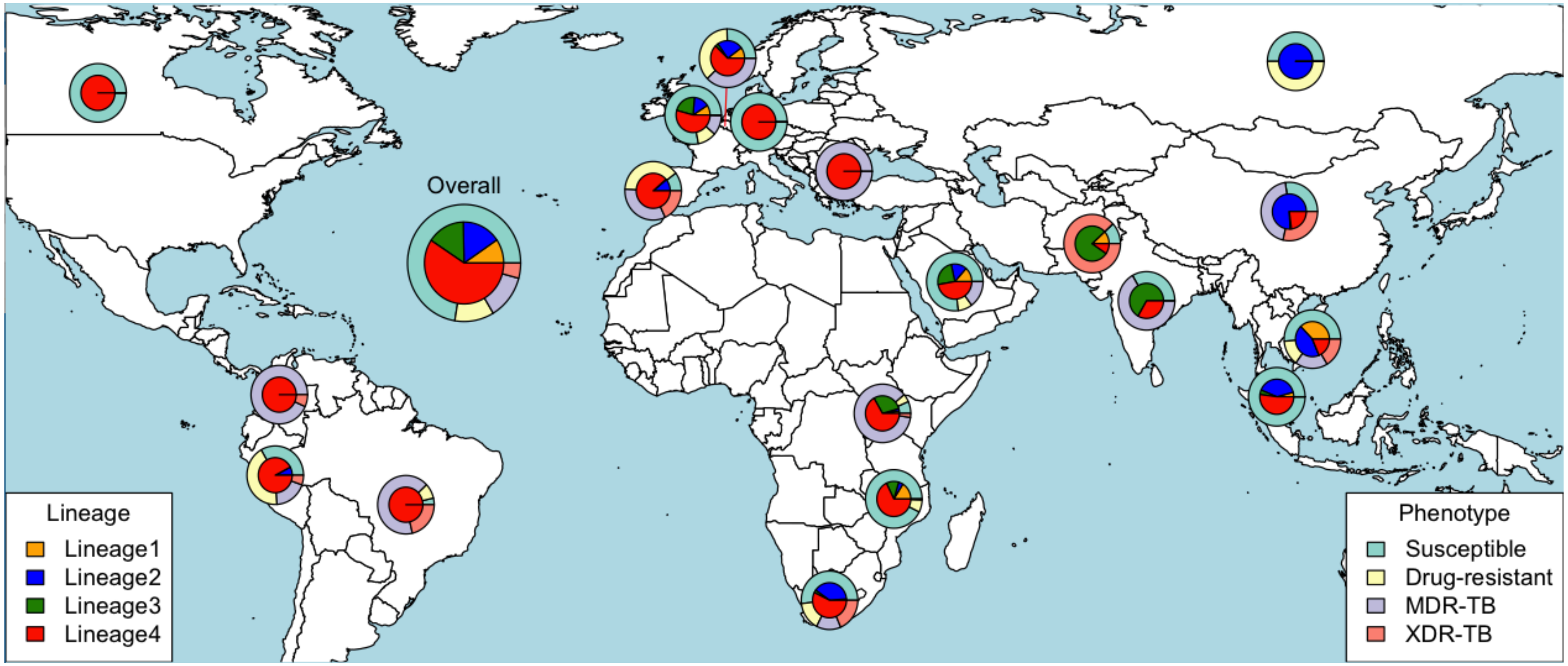


Figure 2. Population structure of 6,465 *M. tuberculosis* isolates based on 102,160 SNPs and 11,122 insertions and deletions spanning the whole genome

Maximum likelihood phylogenetic tree constructed rooted on *M. canetti (not displayed)*, colour-coded by lineage (inner circle) and drug resistance status (outer circle). 'Susceptible' refers to isolates being susceptible to all drugs tested. 'Drug-resistant' refers to strains being resistant to multiple drugs but not classified as multidrug-resistant (MDR-TB) or extensively drug-resistant XDR-TB.

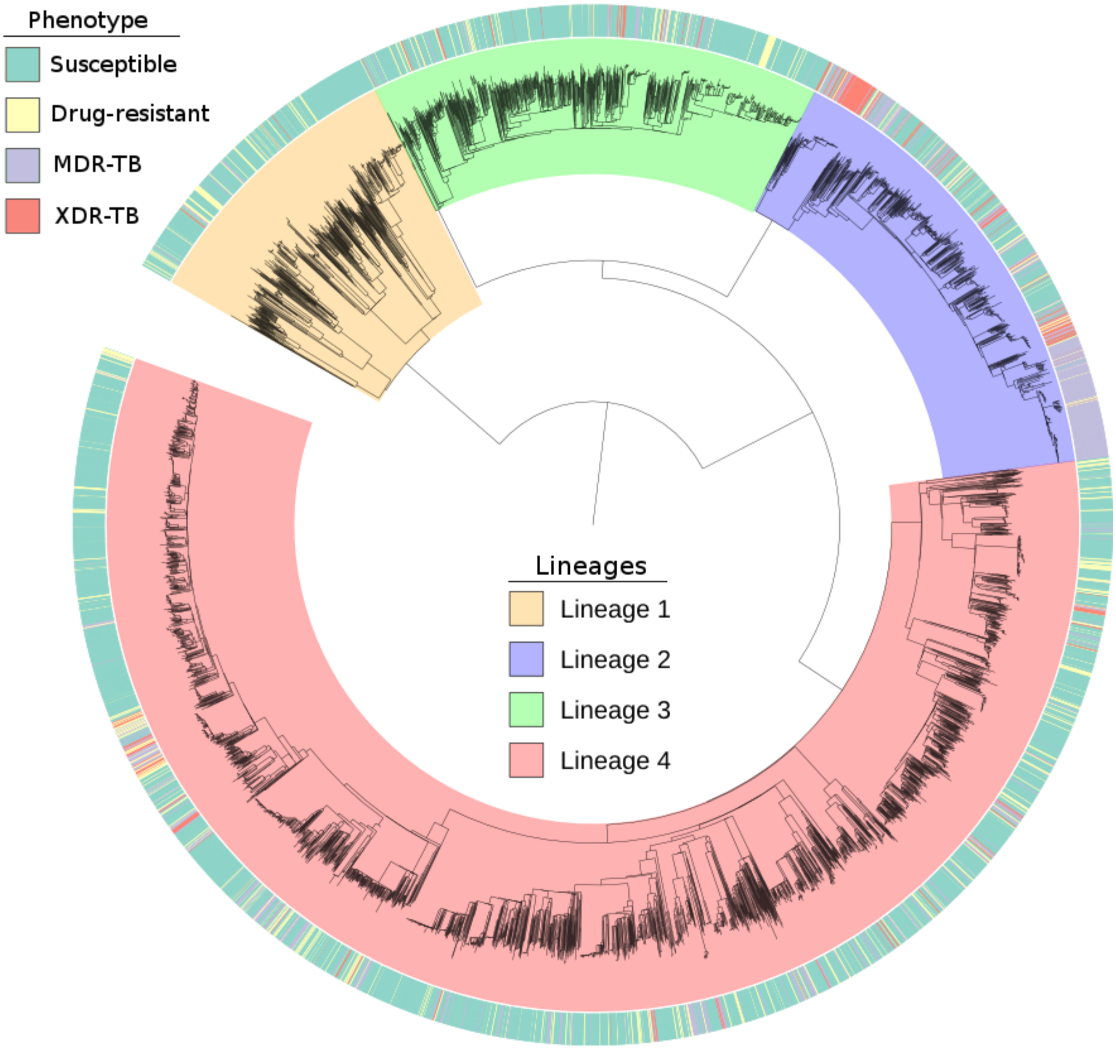


Figure 3. Odds ratios from SNP-drug resistance associations are a potential surrogate for resistance level

Within each drug, boxplots for the log odds ratios (P-values < 10^{-5}) for each gene are arranged by increasing median values (as indicated by the horizontal line in the boxes) to show their relative effect on resistance. Boxplots are colour-coded in blue or red to show whether genes are known to confer ‘low’ or ‘high’ levels of resistance, respectively[20,69–71].

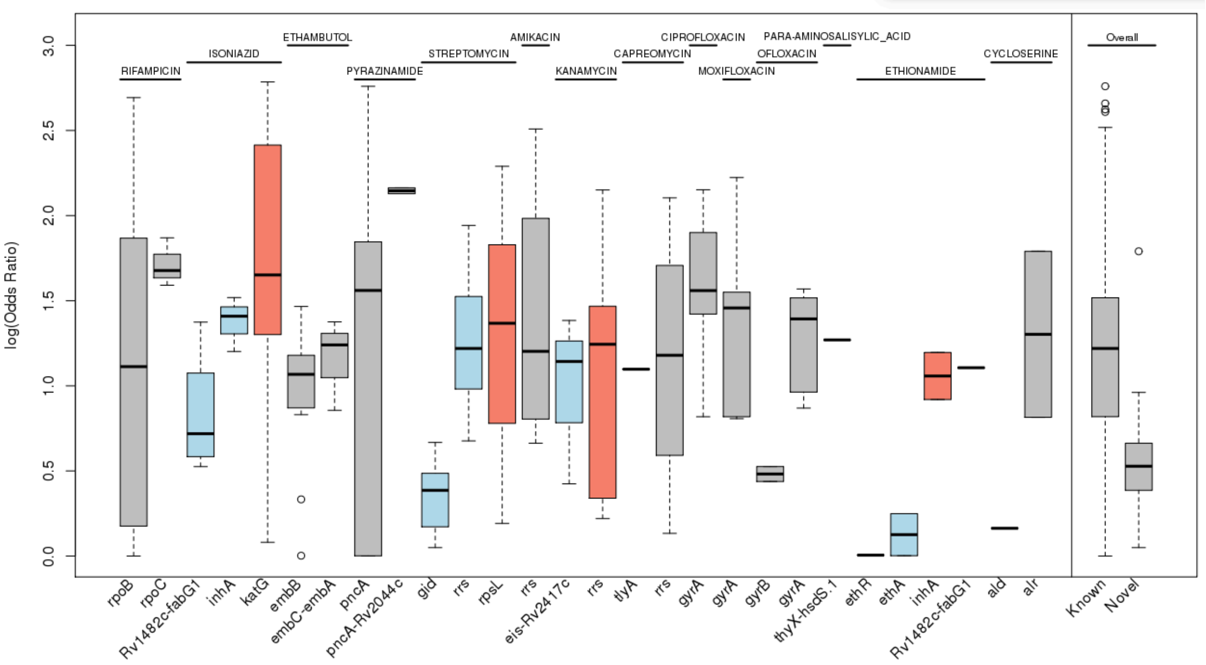


Table 1
MDR-TB and XDR-TB gene-based associations

Comparison	Rv	Gene	Gene-based P-value	NS SNPs	Indels (frame.)	Assoc. SNPs	PhyC SNPs
MDR-TB vs. Susc.	*Rv0667*	*rpoB*	1.98E-139	159	7 (0)	6	8
MDR-TB vs. Susc.	*Rv1908c*	*katG*	2.72E-110	177	12 (9)	1	1
MDR-TB vs. Susc.	*Rv1482c-Rv1483*	*Rv1482c-fabG1*	1.18E-25	8	0	1	1
MDR-TB vs. Susc.	*Rv3795*	*embB*	1.23E-18	168	2 (0)	1	9
MDR-TB vs. Susc.	*Rv1484*	*inhA*	3.13E-18	9	0	2	0
MDR-TB vs. Susc.	*Rv3793-Rv3794*	*embC-embA*	1.85E-13	6	6	1	3
MDR-TB vs. Susc.	*Rv0682*	*rpsL*	2.96E-13	6	0	0	2
MDR-TB vs. Susc.	*Rv3919c*	*gid*	5.22E-11	137	26 (26)	0	1
MDR-TB vs. Susc.	*Rv2427A-Rv2428*	*oxyR'-ahpC*	2.51E-10	17	3	0	3
MDR-TB vs. Susc.	*Rv0721*	*rpsE*	8.10E-08	24	0	0	0
MDR-TB vs. Susc.	*Rv2043c*	*pncA*	1.32E-06	117	25 (22)	0	1
XDR- vs. MDR-TB	*Rv0006*	*gyrA*	5.10E-30	147	0	2	4
XDR- vs. MDR-TB	*rrs*	*rrs*	5.30E-06	91	4	1	2
XDR-TB vs. Susc.	*Rv0667*	*rpoB*	3.04E-203	159	7 (0)	7	5
XDR-TB vs. Susc.	*Rv2043c*	*pncA*	4.52E-143	117	25 (22)	2	0
XDR-TB vs. Susc.	*Rv3795*	*embB*	6.17E-85	168	2 (0)	4	4
XDR-TB vs. Susc.	*Rv1908c*	*katG*	7.38E-83	177	12 (9)	1	1
XDR-TB vs. Susc.	*Rv1482c-Rv1483*	*Rv1482c-fabG1*	3.75E-52	8	0	2	2

Table 1 - continued

Comparison	Rv	Gene	Gene-based P-value	NS SNPs	Indels (frame.)	Assoc. SNPs	PhyC SNPs
XDR-TB vs. Susc.	*Rv3793-Rv3794*	*embC-embA*	2.85E-49	6	6	2	2
XDR-TB vs. Susc.	*Rv0682*	*rpsL*	1.05E-40	6	0	1	2
XDR-TB vs. Susc.	*rrs*	*rrs*	4.66E-29	91	4	2	3
XDR-TB vs. Susc.	*Rv1144-Rv1145*	*Rv1144-mmpL13a*	6.70E-08	33	4	1	0
XDR-TB vs. Susc.	*Rv1484*	*inhA*	6.10E-29	9	0	2	0
XDR-TB vs. Susc.	*Rv0006*	*gyrA*	1.27E-25	147	0	4	4
XDR-TB vs. Susc.	*Rv0668*	*rpoC*	9.57E-19	153	1 (0)	2	0
XDR-TB vs. Susc.	*Rv2427A-Rv2428*	*oxyR'-ahpC*	7.20E-15	17	3	0	0
XDR-TB vs. Susc.	*Rv2936*	*drrA*	1.46E-09	19	0	1	0
XDR-TB vs. Susc.	*Rv3854c*	*ethA*	2.04E-11	163	38 (35)	0	0
XDR-TB vs. Susc.	*Rv3854c-Rv3855*	*ethA-ethR*	5.87E-06	12	0	1	0
XDR-TB vs. Susc.	*Rv2416c-Rv2417c*	*eis-Rv2417c*	5.88E-06	12	1	0	1
XDR-TB vs. Susc.	*Rv3144c-Rv3145*	*PPE52-nuoA*	8.54E-06	24	1	0	0

This table shows loci (protein and RNA coding regions, intergenic regions) associated with MDR- and XDR-TB resistance (P-value < 1x10^{-5}). The column labelled as ‘NS SNPs’ shows the number of non-synonymous SNPs in the genes; the column ‘Indels (frame.)’ refers to the number of small indels resulting in frameshifts in the genes; ‘Assoc. SNPs’ refers to the number of SNPs identified by GWAS and ‘PhyC SNPs’ is the number of homoplastic SNPs identified using the PhyC test. The PhyC test additionally detected *folC*, *pncA-Rv2044c* and *whiB6-Rv3863* loci when comparing MDR-TB against the susceptible group; and *gyrB and thyX-hsdS.1* loci when comparing XDR-TB against susceptible). Similarly, the GWAS using SNPs additionally identified the *ubiA* gene for XDR-TB vs. MDR-TB (2 SNPs) and XDR-TB vs. susceptible.

Table 2
Individual drug gene-based associations in the complete dataset

Drug*	Rv	Gene	Gene-based P-value	NS SNPs	Indels (frame.)	Assoc. SNPs	PhyC SNPs
Isoniazid	*Rv1908c*	*katG*	6.40E-114	177	12 (9)	1	3
Isoniazid	*Rv1482c-Rv1483*	*Rv1482c-fabG1*	8.01E-62	8	0	2	2
Isoniazid	*Rv2427A-Rv2428*	*oxyR'-ahpC*	3.48E-28	17	3	0	3
Isoniazid	*Rv1484*	*inhA*	1.44E-07	9	0	1	1
Rifampicin	*Rv0667*	*rpoB*	2.87E-245	159	7 (0)	6	9
Rifampicin	*Rv0668*	*rpoC*	2.65E-08	153	1 (0)	0	9
Ethambutol	*Rv3795*	*embB*	4.67E-115	168	2 (0)	4	10
Ethambutol	*Rv3793-Rv3794*	*embC-embA*	1.62E-44	6	6	2	5
Ethambutol	*Rv2820c*	.	1.30E-10	16	0	1	0
Ethambutol	*Rv3806c*	*ubiA*	1.36E-10	47	0	0	2
Ethambutol	*Rv3300c*	.	8.02E-08	39	5 (3)	0	0
Ethionamide	*Rv1482c-Rv1483*	*Rv1482c-fabG1*	4.78E-11	8	0	1	2
Ethionamide	*Rv1484*	*inhA*	7.60E-07	9	0	1	0
Pyrazinamide	*Rv2043c*	*pncA*	3.18E-110	117	25 (22)	2	1
Pyrazinamide	*Rv2043c-Rv2044c*	*pncA-Rv2044c*	7.74E-29	4	1	1	1
Streptomycin	*Rv0682*	*rpsL*	1.57E-82	6	0	2	2
Streptomycin	*Rv3919c*	*gid*	1.51E-26	137	26 (26)	0	1
Streptomycin	*rrs*	*rrs*	4.40E-11	91	4	1	3

Table 2 - continued

Drug*	Rv	Gene	Gene-based P-value	NS SNPs	Indels (frame.)	Assoc. SNPs	PhyC SNPs
Amikacin	*rrs*	*rrs*	2.68E-46	91	4	1	1
Kanamycin	*rrs*	*rrs*	7.42E-38	91	4	2	2
Kanamycin	*Rv2416c-Rv2417c*	*eis-Rv2417c*	3.53E-18	12	1	1	1
Capreomycin	*rrs*	*rrs*	2.12E-37	91	4	1	1
Capreomycin	*Rv2172c-Rv2173*	*Rv2172c-idsA2*	2.93E-06	18	0	2	0
Ciprofloxacin	*Rv0006*	*gyrA*	9.30E-43	147	0	2	2
Moxifloxacin	*Rv0006*	*gyrA*	3.51E-22	147	0	2	5
Ofloxacin	*Rv0006*	*gyrA*	3.88E-49	147	0	3	6
D-Cycloserine	*Rv3423c*	*alr*	1.26E-13	57	0	1	0
D-Cycloserine	*Rv0342*	*iniA*	3.37E-08	76	13 (12)	1	0
PAS	*Rv2764c*	*thyA*	3.74E-10	36	4 (4)	0	0
PAS	*Rv2754c-Rv2755c*	*thyX-hsdS.1*	4.27E-07	21	0	1	1

This table shows loci (protein and RNA coding and intergenic regions) associated with resistance to individual drugs (P-value < 1x10^{-5}). The column labelled as 'NS SNPs' show the number of non-synonymous SNPs in the genes; the column 'Indels (frame.)' refers to the number of small indels resulting in frameshifts in the genes; 'Assoc. SNPs' is the number of SNPs identified by GWAS, and 'PhyC SNPs' refers to the number of homoplastic SNPs identified using the PhyC test. * The PhyC test additionally detected other associated loci for Amikacin (*eis-Rv2417c*), Capreomycin and D-Cycloserine (*lhr),* Kanamycin (*thyX-hsdS.1*), Rifampicin (*rpoA*); PAS, Para-aminosalicylic acid.

Table 3
Impact on drug resistance prediction (%) from GWAS findings

	TBDR panel		+ SNPs		+ small indels + SNPs		+ big deletions + small indels + SNPs	
Drug	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
Isoniazid	88	97	88	97	**89**	97	89	97
Rifampicin	91	98	91	98	**92**	98	92	98
Ethambutol	88	92	88	92	88	92	88	92
Ethionamide	75	75	75	73	**82**	70	**86**	70
Pyrazinamide	56	98	56	98	**59**	98	**62**	98
Streptomycin	75	93	76	93	**79**	91	79	91
Amikacin	83	96	83	96	**85**	93	85	93
Kanamycin	86	98	86	98	86	98	86	98
Capreomycin	73	96	73	96	**80**	95	80	95
Ciprofloxacin	88	98	88	98	88	98	88	98
Moxifloxacin	84	90	84	90	84	90	84	90
Ofloxacin	83	93	83	93	83	93	83	93
D-Cycloserine	-	-	**55**	**92**	**61**	90	61	90
PAS	10	100	**20**	99	**40**	94	**65**	94
MDR-TB	88	99	88	99	88	99	**89**	99
XDR-TB	74	96	74	96	74	96	**76**	96

This table shows the sensitivity and specificity achieved by known drug resistance SNPs and indels (TBDR, tbdr.lshtm.ac.uk)[9, 31] when predicting phenotypic drug resistance (“TBDR panel" columns). The SNPs in the TBDR contribute 100% to the stated sensitivity, except rifampicin (99.8%) and ethionamide (99.3%). The other columns show the improvements achieved when including the SNPs, small indels and large deletions found associated with drug resistance in this study. The improvements in sensitivity are highlighted in grey.

Abbreviations: MDR-TB, multidrug-resistant; PAS Para-aminosalicylic acid; Sens., sensitivity; Spec., specificity; SNPs, single nucleotide polymorphisms; XDR-TB, extensively drug-resistant.

Supplementary table 1
Populations contributing to the analysis

Population	N	lineage 1	lineage2	lineage3	lineage4	Susc.	DR	MDR-TB	XDR-TB	ENA Accession
Canada	11	0	0	0	11	11	0	0	0	SRA020129
Brazil	108	0	0	0	108	4	9	72	23	**PRJEB10385**
Colombia	15	0	0	0	15	0	0	14	1	**PRJEB10385**
Peru	78	0	6	0	72	25	32	17	4	**PRJEB10385**
Bulgaria	2	0	0	0	2	0	0	2	0	**PRJEB10385**
Germany	20	0	0	0	20	20	0	0	0	ERP006619
Portugal	183	0	20	1	162	19	71	60	33	**ERP002611**
Russia	2	0	2	0	0	1	1	0	0	ERP00192
China	161	0	122	2	37	44	0	71	46	SRP018402
Vietnam	43	16	19	0	8	22	6	8	7	**PRJEB10385**
India	3	0	0	2	1	1	0	2	0	**PRJEB10385**
Pakistan	42	5	0	33	4	5	0	0	37	**ERP008770**
Saudi Arabia	74	10	11	18	35	57	6	11	0	**PRJEB10385**
Malawi	1646	264	71	195	1116	1526	112	8	0	**ERP000436**
South Africa	594	8	231	15	340	308	93	83	110	**PRJEB10385**
Uganda	45	1	1	13	30	3	2	39	1	**ERP000520**
WHO*	138	14	34	4	86	35	51	52	0	**ERP013054**
Mixed**	96	4	38	4	50	96	0	0	0	ERP001037
UK	3204	295	466	706	1737	2500	343	351	10	ERX511672

Population	N	lineage 1	lineage2	lineage3	lineage4	Susc.	DR	MDR-TB	XDR-TB	ENA Accession
Total	6465	617	1021	993	3834	4677	726	790	272	
%	100	9.5	15.8	15.4	59.3	72.3	11.2	12.2	4.2	

Susc. = susceptible; DR = resistant to at least one drug but not MDR-TB/XDR-TB ; *Bangladesh (8), China (1), Nepal (4), Pakistan (1), Philippines (4), South Korea (39), Thailand (1), Cameroon (1), Central African Republic (1), *Equatorial Guinea (1),* Guinea (1), Morocco (4), Niger (1), Nigeria (1), Democratic Republic of Congo (4), Rwanda (15), Gemany (12), Kazakstan (1), Portugal (1), Spain (2), Brazil (7), Columbia (1), Domican Republic (1), Peru (31); ** Malaysia, South Africa, and Thailand (96); *** PRJNA183624, PRJNA 235615, PRJEB10385; **bolded ENA accession numbers** include sequencing performed as part of the TB Global Drug Resistance Collaboration (http://pathogenseq.lshtm.ac.uk/#tuberculosis).

Supplementary table 2
Drugs susceptibility test data (resistant/tested) and the phenotypes considered.

Drug	Lineage 1	Lineage 2	Lineage 3	Lineage 4	Total Resistant	(%)
Rifampicin (RIF)	26/609	549/928	78/985	488/3529	1141	-18.9
Isoniazid (INH)	87/608	569/938	157/985	723/3546	1536	-25.3
Ethambutol (EMB)	16/403	357/858	36/839	236/2707	645	-13.4
Pyrazinamide (PZA)	20/393	261/638	39/796	164/2259	484	-11.9
Streptomycin (STR)	23/227	450/718	44/293	369/1957	886	-27.7
Capreomycin (CAP)	Jan-15	125/347	Oct-63	91/579	227	-22.6
Amikacin (AMK)	May-16	128/254	28/70	66/546	227	-25.6
Kanamycin (KAN)	May-17	128/320	28/63	88/506	249	-27.5
Moxifloxacin (MOX)	0/15	66/232	Feb-38	20/351	88	-13.8
Ofloxacin (OFL)	01-Feb	150/281	Jan-22	135/388	287	-41.4
Ethionamide (ETH)	01-Jun	102/273	Feb-34	117/284	222	-37.2
Ciprofloxacin (CIP)	May-41	Feb-24	32/101	20/160	59	-18.1
PAS	0/0	7/119	0/0	13/136	20	-7.8
D-Cycloserine (CYS)	0/0	39/117	0/0	17/131	56	-22.6

Supplementary Table 2 - continued

Phenotype	**Lineage 1** **N (%)**	**Lineage 2** **N (%)**	**Lineage 3** **N (%)**	**Lineage 4** **N (%)**	**Total**	**(%)**
Susceptible	516 (83.6)	408 (40.0)	819 (82.5)	2935 (76.6)	4678	-72.4
Drug resistant	77 (12.5)	78 (7.6)	102 (10.3)	407 (10.6)	664	-10.3
MDR-TB	18 (2.9)	393 (38.5)	43 (4.3)	380 (9.9)	834	-12.9
XDR-TB	6 (1.0)	142 (13.9)	29 (2.9)	112 (2.9)	289	-4.5
Total	617	1021	993	3834	6465	
(%)	-9.5	-15.8	-15.4	-59.3	-100	

Drug resistant = Resistant to at least 1 drug but not MDR-TB/XDR-TB; MDR-TB, multidrug-resistant; XDR-TB, extensive drug-resistant; MOX and OFL are Fluoroquinolones (FLQ); CAP, KAN and AMK are second-line injectables; PAS Para-aminosalicylic acid

Supplementary table 3
Allele frequency of resistance mutations

Gene	Mutation	Susceptible %	DR %	MDR-TB allele frequency %	XDR-TB allele frequency %
katG	S315T	0.4	40.5	79.8	78
rpoB	S450L	0.1	25.6	67	52
rpsL	K43R	0.5	16	45.8	25.6
embB	M306V	0	10.5	34.9	39.4
rrs	A1401G	0	7.4	11.7	63.4
Rv1482c-fabG1	C-15T	0.3	25	11.5	31.9
embB	M306I	0.3	7.2	22.1	37.4
gyrA	A90V	0	2.5	3.1	32.7
rrs	A514C	0.2	3.2	5.7	27.2
gyrA	D94G	0.3	5.6	2	27.6
gid	L79S	0	1.2	2.6	22
rpoB	L452P	0	1.4	4.5	19.7
ethA-ethR	T-65C	0	0.7	5.4	18.9
Rv1482c-fabG1	T-8A	0	1.2	3.7	19.3
rpoB	D435V	0	2.5	6.3	13.8
rpoB	D435G	0	0.3	0.3	19.3
ubiA	V188A	0	0.1	0.3	18.5
rpoB	I1106T	0	0.2	0	18.5
inhA	S94A	0	7.6	1.2	7.5
Rv1482c-fabG1	G-17T	0	0.8	1.3	11.4

Supplementary table 3 - continued

Gene	**Mutation**	**Susceptible %**	**DR**	**MDR-TB allele frequency %**	**XDR-TB allele frequency %**
inhA	I194T	0	4	1.9	7.5
ubiA	A249T	0	0.6	0.9	10.6
PPE52-nuoA	G-314T	0	0.7	1.2	10.2
eis-Rv2417c	C-10T	0	2.5	3.9	5.5
rpsL	K88R	0.1	3.8	5.4	2.4
iniA	H42R	0	0.6	0.7	10.2
gyrA	D94A	0	1.9	1.3	8.3
alr	L113R	0	0.7	0.7	9.8
pncA	Q10*	0	1.1	7.3	2.8
embB	Q497R	0	1.7	6.1	2.4
rpoB	D435Y	0	1.5	3.2	4.7
rpoB	H445Y	0	2.7	3.7	2
gyrA	S91P	0.1	1.8	1.3	5.1
embC-embA	C-12T	0	0.4	3.7	3.9
embB	G406A	0	0.8	3.4	3.5
pncA	Q10P	0	0.3	5.8	1.6
eis-Rv2417c	G-12A	0	0.3	6	0.8
embC-embA	C-16T	0	1.8	2	3.1
embC-embA	C-16G	0	0.6	1.6	4.3
rpoB	H445D	0	2.6	2.9	0.8
gyrA	D94Y	0	0.6	0.6	5.1

Supplementary table 3 - continued

Gene	Mutation	Susceptible %	DR	MDR-TB allele frequency %	XDR-TB allele frequency %
thyX-hsdS.1	G-16A	0	1.4	1.8	2.8
rpoB	L731P	0	2.1	1	2.8
embB	G406D	0	1.7	3.2	0.8
pncA	V125G	0	2	1.3	2.4
embC-embA	C-11A	0	1.3	0.7	3.5
katG	S315R	0	1.7	0.6	3.1
pncA-Rv2044c	T-11C	0	0.7	3.4	1.2
katG	S315N	0	1.7	1.9	1.6
gyrA	D94N	0.1	0.8	0.3	3.9
embB	M423T	0	2	0.7	2.4
gid	A80P	0	2	0.7	2.4
embC-embA	G-43C	0	0.3	2	2.4
embB	D354A	0	0.7	2.3	1.6
Rv2172c-idsA2	A-65G	0	1.3	0.6	2.8
embC-embA	C-12A	0	1.3	0.6	2.8
embB	P397T	0	1.3	0.6	2.8
rrs	C517T	0	0.8	1.9	1.6
eis-Rv2417c	G-14A	0	0.3	1.2	2.8
embB	G406S	0	0.7	2.2	1.2
rpoB	H445R	0.1	0.5	1.9	1.6
embB	D1024N	0	0.7	1.5	1.6

Supplementary table 3 - continued

Gene	**Mutation**	**Susceptible %**	**DR**	**MDR-TB allele frequency %**	**XDR-TB allele frequency %**
oxyR'-ahpC	G-48A	0	0.7	0.7	2.4
alr	M343T	0	0.9	0.4	2.4
rpoB	S450W	0	0.8	1.9	0.4
oxyR'-ahpC	C-52T	0.1	0.6	1.6	0.8
rpoB	H445L	0	0.7	1.2	1.2
pncA	V139M	0	0.4	0.1	2.4
rpoB	H445N	0.1	2.1	0.4	0
rpoB	L430P	0.1	1.1	1.3	0
embC-embA	C-8T	0	0.2	0.7	1.6
rpoB	I491F	0.2	2.2	0	0
pncA	W68*	0	0.2	1.5	0.8
Rv1482c-fabG1	T-8C	0.4	0.4	0.9	0.8
pncA	Q141P	0	0.7	0.9	0.8
gyrA	D94H	0	0.2	0.7	1.2
rrs	A514T	0	1.3	0.3	0.4
rpoB	M434I	0	0.1	0.3	1.6

DR = Resistant to at least 1 drug but not MDR-TB/XDR-TB; multidrug-resistant; XDR-TB, extensively drug-resistant

Supplementary table 4
SNP-based GWAS results in each lineage

Lineage	Gene	Position	Drug	Min P-value	Susc.	DR	MDR-TB	XDR-TB
4	gyrA	7570	X v M or SUS	1.51E-15	0.001	0.024	0.032	0.329
4	gyrA	7572	X v SUS	8.92E-21	0.001	0.018	0.013	0.051
3,4	gyrA	7581	X v SUS	1.17E-21	0.001	0.016	0.016	0.103
4	gyrA	7582	KAN	4.40E-09	0.004	0.076	0.034	0.359
2	gyrA	7582	X v M	1.70E-08	0.004	0.076	0.034	0.359
4	gyrA	7582	X v M or SUS	8.52E-07	0.004	0.076	0.034	0.359
2	rpoB	760314	M v SUS	4.92E-22	0	0.004	0.006	0.004
3	rpoB	761108	X v SUS	3.44E-14	0	0.001	0.003	0.016
2-4	rpoB	761109	M or X v SUS, RMP	3.34E-28	0	0.015	0.032	0.048
3,4	rpoB	761110	X v M, X or M v SUS, RMP	3.35E-85	0	0.029	0.066	0.337
1,2,4	rpoB	761139	X or M v SUS	3.46E-16	0.001	0.076	0.071	0.028
1-4	rpoB	761139	RMP	1.61E-97	0.001	0.076	0.071	0.028
1,2,4	rpoB	761140	M or X v SUS, RMP	2.66E-17	0.001	0.01	0.034	0.028
1-4	rpoB	761155	M or X v SUS, RMP	1.17E-219	0.001	0.267	0.695	0.524
2,4	rpoB	761161	M or X v SUS	9.67E-18	0	0.013	0.044	0.197
4	rpoB	763123	X v M or SUS	1.13E-17	0	0.002	0	0.185
3	rpoC	764666	X or M v SUS, RMP	3.74E-29	0	0.003	0.004	0.016
2	rpoC	764819	M v SUS	3.33E-18	0	0.002	0.013	0

Supplementary table 4 - continued

Lineage	Gene	Position	Drug	Min P-value	Susc.	DR	MDR-TB	XDR-TB
4	rpoC	766823	X v SUS	1.64E-06	0	0.013	0.006	0.028
1	rpoC	767123	MDR or XDR v SUS, RMP	3.91E-24	0	0.003	0.015	0.016
2-4	rpsL	781687	MDR or XDR v SUS, STM	1.65E-45	0.005	0.159	0.458	0.257
2-4	rpsL	781822	STM	4.16E-10	0.002	0.041	0.061	0.024
1	rrs	1472358	STM	5.12E-06	0	0.010	0.001	0
4	rrs	1472359	STM	2.66E-13	0.002	0.047	0.061	0.276
3,1	rrs	1472359	M or X v SUS	5.71E-18	0.002	0.047	0.061	0.276
1	rrs	1472362	M or X v SUS, STM	3.52E-71	0	0.009	0.019	0.016
3	rrs	1472751	X v SUS	2.28E-10	0	0.004	0.006	0.004
2,4	rrs	1473246	AMK, CAP, KAN	6.68E-42	0	0.075	0.120	0.651
3	rrs	1473246	STM	3.05E-09	0	0.075	0.120	0.651
2-4	rrs	1473246	X v SUS or M	7.73E-246	0	0.075	0.120	0.651
1	pncB1	1499617	PZA	3.20E-06	0	0.003	0.003	0.004
1	echA12	1660232	X v SUS	3.91E-24	0	0.003	0.003	0.004
2	Rv1482c-fabG1	1673425	ETH	1.91E-04	0.003	0.251	0.115	0.319
1-4	Rv1482c-fabG1	1673425	M or X v SUS, INH	4.07E-56	0.003	0.251	0.115	0.319
4	Rv1482c-fabG1	1673432	X v SUS	2.63E-15	0.004	0.016	0.050	0.205

Supplementary table 4 - continued

Lineage	Gene	Position	Drug	Min P-value	Susc.	DR	MDR-TB	XDR-TB
4	inhA	1674481	X v SUS	8.54E-46	0	0.076	0.010	0.075
1,4	inhA	1674782	M or X v SUS	3.91E-24	0	0.040	0.019	0.075
1-4	katG	2155168	M or X v SUS, INH	3.26E-286	0.004	0.424	0.820	0.795
4	pncA	2288868	X v SUS, PZA	1.60E-14	0	0.02	0.018	0.024
1	pncA	2288952	M or X v SUS, PZA	3.91E-24	0	0.009	0.003	0.004
2	pncA-Rv2044c	2289252	PZA	1.12E-08	0	0.008	0.036	0.016
4	eis-Rv2417c	2715342	KAN	1.93E-08	0	0.025	0.043	0.056
2	oxyR'-ahpC	2726141	X v SUS	4.53E-08	0.001	0.008	0.022	0.008
2	alr	3841083	Cycloserine	1.67E-08	0	0.007	0.008	0.100
2,4	embC-embA	4243217	X or M v SUS, EMB	2.09E-14	0	0.026	0.043	0.079
3,4	embC-embA	4243221	X v SUS, EMB	1.70E-32	0	0.017	0.043	0.067
4	embC-embA	4243222	X v SUS, EMB	2.62E-10	0	0.013	0.01	0.036
1-4	embB	4247429	M or X v SUS, EMB	1.28E-47	0.001	0.109	0.358	0.399
1-4	embB	4247431	M or X v SUS, EMB	1.58E-51	0.003	0.071	0.221	0.375
1	embB	4247574	X or M v SUS, EMB	3.63E-07	0	0.007	0.023	0.016
4	embB	4247702	X v SUS, EMB	2.62E-08	0	0.013	0.006	0.028
4	embB	4247729	X or M v SUS, EMB	8.81E-13	0	0.008	0.023	0.012
2,4	embB	4247730	X v SUS, EMB	3.31E-12	0	0.025	0.066	0.043
4	embB	4247781	X v SUS	9.20E-10	0	0.02	0.007	0.024
1,3,4	embB	4248003	M or X v SUS, EMB	1.33E-26	0	0.021	0.065	0.031

Supplementary table 4 - continued

Lineage	Gene	Position	Drug	Min P-value	Susc.	DR	MDR-TB	XDR-TB
3	embB	4249583	X v SUS, EMB	5.84E-23	0	0.007	0.015	0.016
4	ubiA	4269271	X v M or SUS	1.01E-16	0	0.001	0.003	0.185
3	ethA	4326435	X v SUS	3.44E-14	0	0.001	0	0.016
4	ethA-ethR	4327484	X v SUS	9.24E-44	0	0.007	0.054	0.192
4	ethR	4328127	X v SUS	9.30E-10	0	0.020	0.007	0.024
4	gid	4407965	X v SUS	6.27E-10	0	0.020	0.008	0.029

X = XDR-TB, M = MDR-TB, SUS = Pan susceptible, DR = Resistant to at least one drug but not MDR-TB/XDR-TB, RIF = rifampicin, INH = isoniazid, ETH = ethionamide, EMB = ethambutol, KAN - kanamycin

Supplementary table 5

Detected co-occurrence of mutations at drug resistance associated loci (Fisher exact test $P<10^{-8}$)

Drug	Resistance gene	Co-occurring gene	Fisher exact test p-value
Rifampicin	*rpoB*	*rpoC**	< 2.2e-16
Rifampicin	*rpoB*	*rpoA**	6.0e-09
Isoniazid	*katG*	*ahpC**	< 2.2e-16
Pyrazinamide	*pncA*	*pncB2*	1.4e-13
Ethambutol	*embB*	*ubiA*	< 2.2e-16
PAS	*thyA*	*thyX-hsdS.1*	< 2.2e-16

PAS = Para-aminosalicylic acid; underlying overall and lineage data are presented in **Supplementary table 6**; * known compensatory mechanisms

Supplementary table 6
Co-occurrence of mutations at drug resistance associated loci with a breakdown by lineage

		rpoB (81-bp rifampin resistance-determining region)									
		Overall		Lineage 1		Lineage 2		Lineage 3		Lineage 4	
		WT	Mut.	WT	Mut.	WT	Mut.	WT	Mut.	WT	Mut.
rpoC	WT	4964	873	557	30	417	322	888	49	3102	472
	Mut.	138	477	15	15	25	251	28	28	70	183
rpoA	WT	5060	1308	564	45	439	553	915	76	3142	634
	Mut.	43	42	8	0	3	20	1	1	31	21

		katG									
		Overall		Lineage 1		Lineage 2		Lineage 3		Lineage 4	
		WT	Mut.	WT	Mut.	WT	Mut.	WT	Mut.	WT	Mut.
ahpC promoter	WT	4959	1390	554	58	472	525	826	156	3107	651
	Mut.	35	62	4	0	5	16	5	5	21	41

		pncA									
		Overall		Lineage 1		Lineage 2		Lineage 3		Lineage 4	
		WT	Mut.	WT	Mut.	WT	Mut.	WT	Mut.	WT	Mut.
pncB2	WT	5608	599	560	23	704	280	953	18	3391	278
	Mut.	116	59	24	0	9	0	13	0	70	59
pncB1	WT	5576	647	528	15	701	280	927	17	3420	335
	Mut.	147	11	58	8	12	0	37	1	40	2

		ethA									
		Overall		Lineage 1		Lineage 2		Lineage 3		Lineage 4	
		WT	Mut.	WT	Mut.	WT	Mut.	WT	Mut.	WT	Mut.
pyrG	WT	5922	285	541	33	914	59	933	43	3534	150
	Mut.	143	15	38	1	7	2	14	0	84	12
Rv0565c	WT	5969	292	562	34	914	56	915	42	3578	160
	Mut.	90	8	17	0	6	5	30	1	37	2

Supplementary table 6 - continued

		embB									
		Overall		Lineage 1		Lineage 2		Lineage 3		Lineage 4	
		WT	Mut.	WT	Mut.	WT	Mut.	WT	Mut.	WT	Mut.
ubiA	WT	5033	1281	502	91	489	475	886	97	3156	618
	Mut.	45	104	21	3	3	54	9	1	12	46

		thyA									
		Overall		Lineage 1		Lineage 2		Lineage 3		Lineage 4	
		WT	Mut.	WT	Mut.	WT	Mut.	WT	Mut.	WT	Mut.
thyX-hsdS1	WT	6332	36	600	4	982	10	973	14	3777	8
	Mut.	67	21	13	0	14	14	6	0	34	7

Each table contains the number of isolates with and without mutations (‘mutant’ (Mut) & ‘wild type’ (WT) respectively) at each pair of drug resistance associated loci effects identified or known compensatory effects. ‘Mutant’ refers to isolates with SNP and indel non-synonymous amino acid changes. Synonymous amino acid changes and deep phylogenetic mutations were discarded. Cells with grey background show statistically significant correlations (Fisher exact test P<0.02), i.e. pairs of genes frequently mutated in the same isolates, whereas white background indicates lack of statistical significance. This analysis points to putative epistatic and compensatory relationships.

Supplementary table 7
Protein structural modelling of *alr* reveals low frequency mutations conferring higher instability

Genomic position	Mutation	Overall Mutation Frequency	Resist. Freq.	mCSM*	DUET*	mCSM-Lig**	Distance from CYS**	mCSM-PPI***
3840259	Y388D	0.0009	0	-3.369	-3.384	-3.737	2.682	-2.819
3840258	Y388C	0.0002	0	-1.889	-1.704	-1.938	2.682	-2.489
3840393	M343T_B	0.0031	0.0358	-2.118	-2.085	0.368	3.636	-0.195
3840708	S238L	0.0002	0	0.611	1.192	0.69	4.246	-0.551
3840952	K157E	0.0003	0	-1.483	-1.455	-1.841	4.474	-0.075
3840636	P262Q	0.0012	0	-2.015	-2.069	0.279	4.987	-0.863
3840717	S235W	0.0002	0	-0.807	-1.46	0.706	5.212	-0.588
3840402	R340L_B	0.0003	0	-0.57	0.616	0.16	5.389	-0.629
3840643	L260V	0.0002	0	-1.244	-1.554	-2.467	6.992	-0.419
3840639	S261N	0.0002	0	-1.443	-1.606	-0.482	7.116	-0.248
3841083	**L113R**	**0.0057**	**0.4461**	**-0.961**	**-0.956**	**-1.721**	**8.477**	**-0.423**

CYS = D-cycloscerine; * protein stability; ** drug binding, *** protein-protein interactions; bolded the mutation that was statistically significant; grey – less stability

We applied four measures to quantify the enthalpic effects (the change in Gibbs free energy - ΔΔG) of point mutations on overall protein structure stability (mCSM and DUET), protein-protein interactions (mCSM-PPI) and interaction with substrate/drug (mCSM-Lig). Negative values indicate a destabilising effect, with the most destabilising highlighted in grey, and positive values indicating an increase in stability. The geometrical distance from the mutation to the drug binding position is also provided. The mutation that was statistically significant with the largest resistance frequency (**L113R**) has a relatively large destabilising effect both on the overall protein structure and in drug binding, yet it is the furthest from the site of drug interaction.

Supplementary table 8
Gene-based small insertion and deletion (indel) associations

Drug	Gene	indels/Kb	Total No. positions	LengthMedian (bp)	Length Range (bp)	Assoc. P-value
MDR-TB vs. Susc.	*pncA*	44.72	25	1	1-15	3.83E-10
MDR-TB vs. Susc.	*rpoB*	2.27	7	6	3-9	3.49E-06
MDR-TB vs. Susc.	*embCAB promoter*	72.29	6	1	1-2	1.71E-04
XDR-TB vs. Susc.	*ethA*	25.89	38	1	1-10	4.25E-54
XDR-TB vs. Susc.	*pncA*	44.72	25	1	1-15	5.51E-38
XDR-TB vs. Susc.	*rpoB*	2.27	7	6	3-9	1.31E-12
XDR-TB vs. Susc.	*embCAB promoter*	72.29	6	1	1-2	1.29E-29
XDR- vs. MDR-TB	*pncA*	44.72	25	1	1-15	1.50E-04
XDR- vs. MDR-TB	*katG*	5.40	12	1.5	1-12	2.33E-02
Isoniazid	*katG*	5.40	12	1.5	1-12	2.82E-05
Rifampicin	*rpoB*	2.27	7	6	3-9	1.25E-10
Ethionamide	*ethA*	25.89	38	1	1-10	7.22E-09
Capreomycin	*tlyA*	3.73	3	2	2-10	1.21E-12
Capreomycin	*rrs*	2.61	4	1	1-1	2.37E-10
Streptomycin	*gid*	35.66	24	1	1-14	1.45E-09
Pyrazinamide	*pncA*	44.72	25	1	1-15	5.27E-38
Cycloserine	*ald*	10.77	12	1	1-5	5.35E-03
Kanamycin	*rrs*	2.61	4	1	1-1	9.29E-05

Susc. = susceptible; MDR-TB = multi-drug resistance; XDR-TB = extensive drug resistance

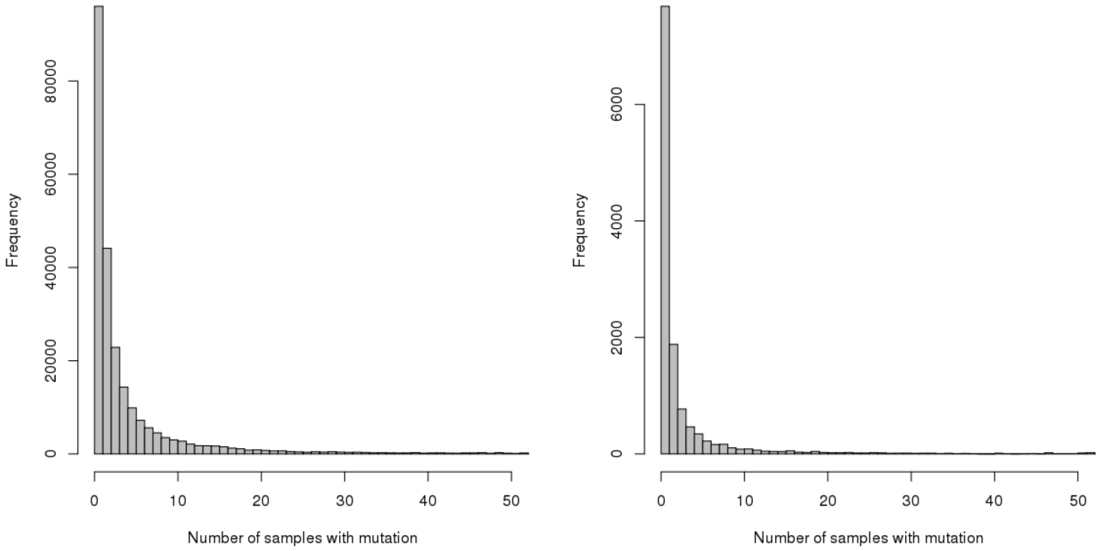
Supplementary table 9
Large deletions in candidate drug resistance regions

Gene	No. samples	Drug	No. DR	No. XDR-TB	mean size (bp)	Size range (bp)
dfrA/thyA	5	PAS	1	3	6,396	2,825-7,912
pncA	12	PZA	1	3	1,402	446-4,670
ethA/ethR	7	ETH	3	3	3,667	1,513-5,271
katG	3	INH	3	0	5,729	4,789-7,608

DR = Resistant to at least one drug but not MDR/XDR-TB; XDR-TB = extensive drug resistance; PAS Para-aminosalicylic acid, ETH Ethionamide, PZA Pyrazinamide, INH Isoniazid

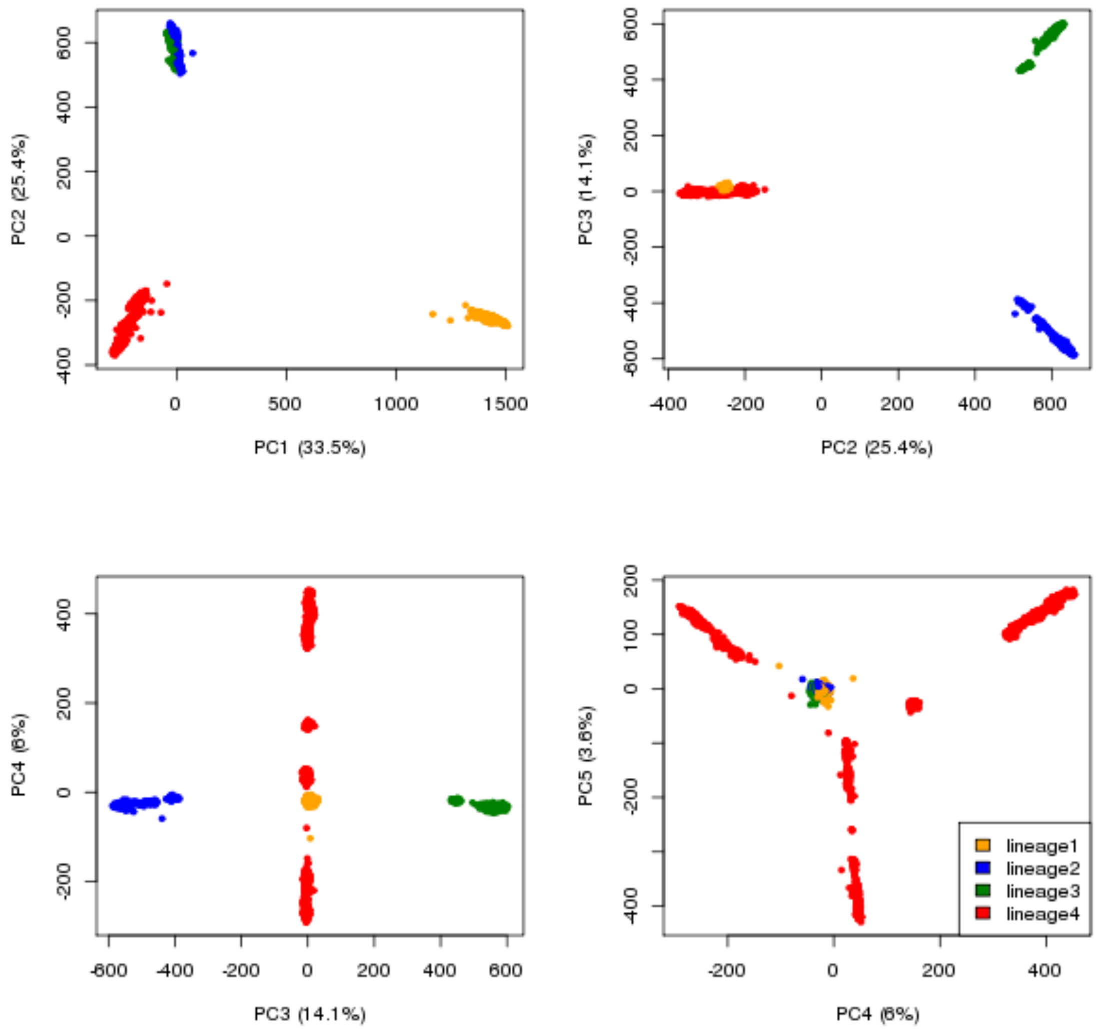
Supplementary figure 1

Allele frequency spectra for SNPs (left) and small insertions and deletions (indels, right)

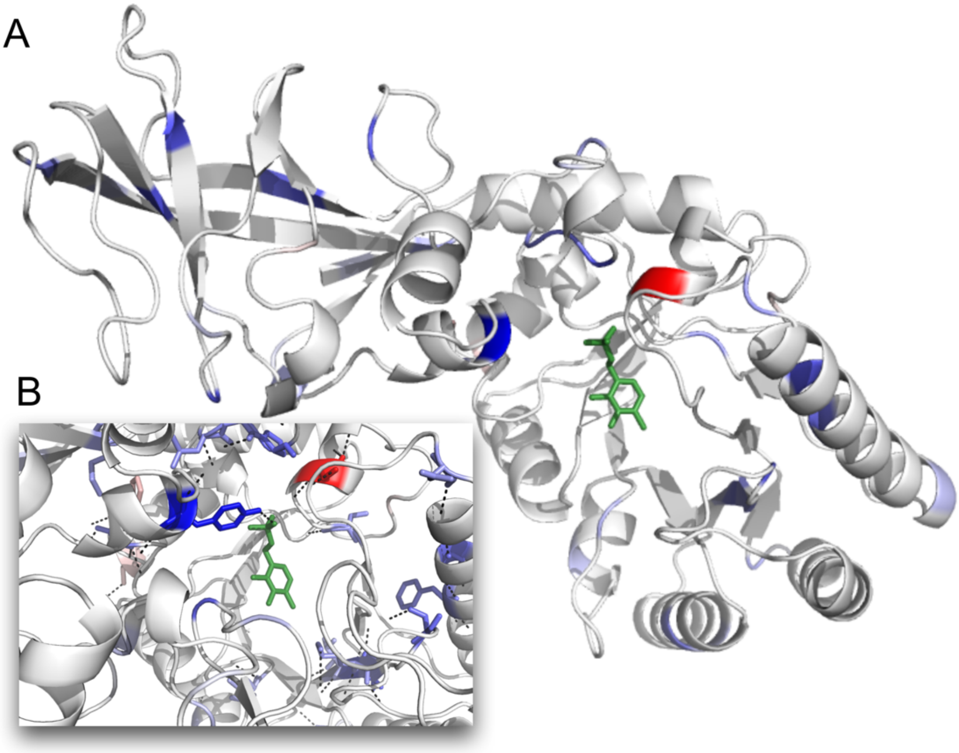


Supplementary figure 2

Principal component (PC) analysis confirms lineage and sub-lineage based population structure (total variation explained across five components is 79%)



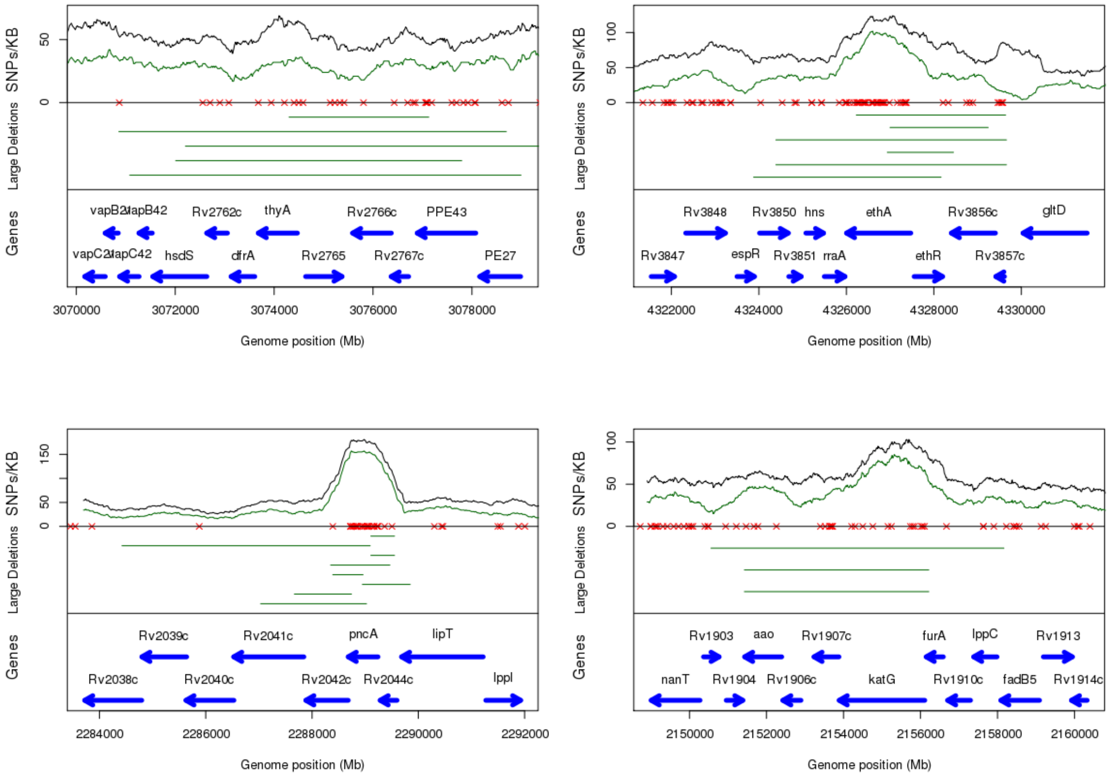
Supplementary figure 3
Protein structure for alr



Alanine racemase mutational map showing position and effect of mutations based on measure of protein stability by DUET. Unfavourable mutations are depicted in blue and favourable mutations in red, where colour intensity reflects extent of effect. The PLP co-factor shown as a stick representation in green. (A) shows the protomer structure of alanine racemase depicted as a cartoon with the PLP co-factor shown as sticks. Insert (B) shows the active site with residues that have been identified in the GWAS depicted as sticks and their hydrogen bonding propensity shown as dashed black lines.

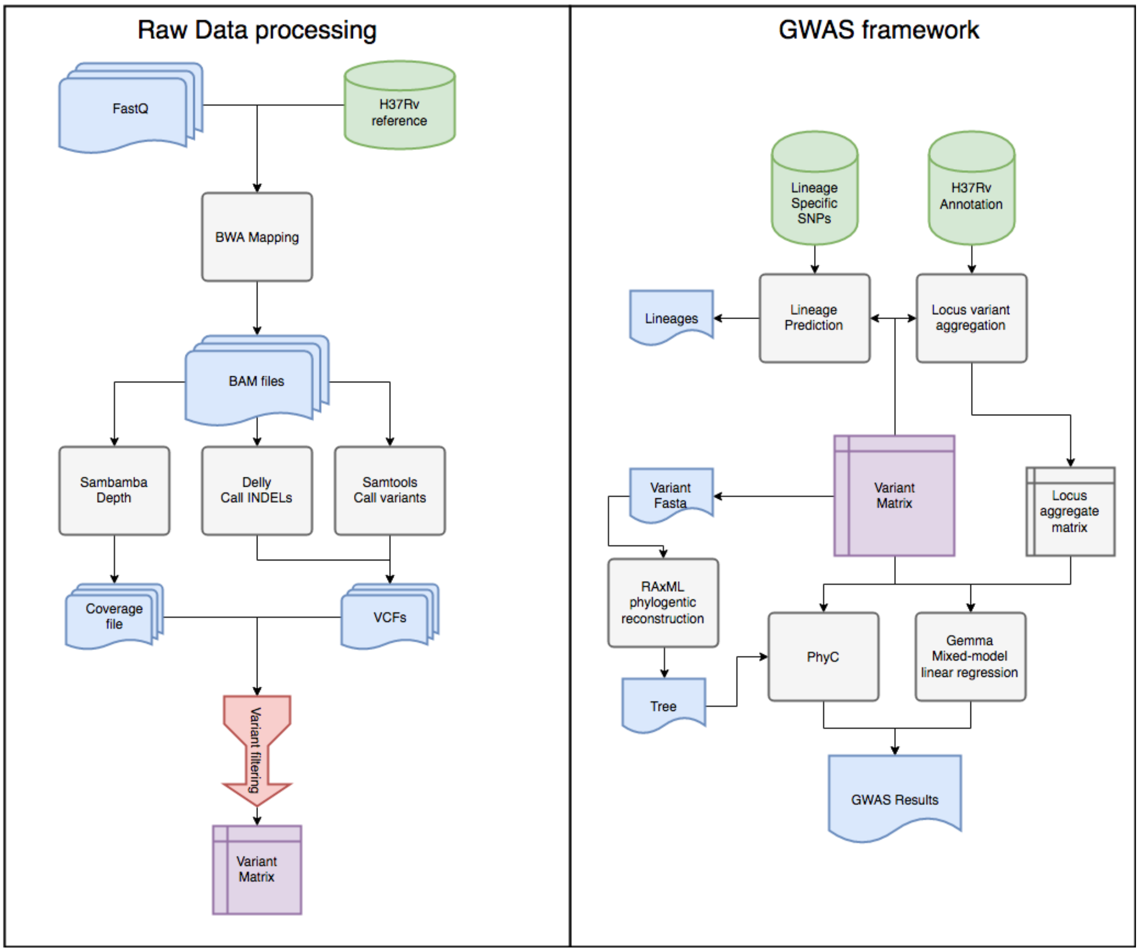
Supplementary figure 4

Polymorphisms in regions surrounding *ethA* (top left), *thyA* (top right), *pncA* (bottom left), and *katG* (bottom right) using the complete dataset (n=6,465)



The top panel in the figures shows the density of SNPs per Kb (green – non-synonymous, black – all). The red crosses show the location of the small indels. The middle panel shows the location of the large deletions found in samples used in this study. The lower panel shows the location of the candidate regions and flanking genes.

Supplementary figure 5
The analytical workflow, including procedures adopted for raw sequence data processing and the genome-wide association study (GWAS) approach



Supplementary data 1

All genes and operons identified as being significant in analyses, with their mutations and indels; Minor allele frequency (<10 samples) and major allele frequency (>=10 samples) mutations are presented separately, and association SNP hits are bolded.

* stop codon

Locus	Low frequency mutations (<10 samples)	High freq. (>= 10 samples)
gyrB	E21K, R40P, N66H, L70F, A78D, I84V, T88N, D97E, A130S, E145K, Q148E, K159R, A162S, K165N, F180V, T183M, L204M, D210Y, D225E, A242P, K247N, H252Y, H263N, T267S, K268R, I271M, V276A, S279C, G280D, G282D, E299Q, A323G, D334G, D340G, D342E, T346S, A355T, A355S, K361T, T393S, P400R, A416V, R421H, V427M, K430Q, A432V, T433I, D434G, D434E, A443T, C445S, R446H, S447Y, S447F, P450S, R451H, L455V, V457I, **D461N**, **D461H**, G464S, A467V, K498N, N499D, N499T, N499S, N499K, T500P, **T500N**, **T500I**, A504T, **A504V**, I505V, I506S, G512R, I519V, G520A, K521E, K526Q, H560Y, H560R, V561A, R575H, S576G, D584H, A593V, K596N, G606A, M616V, E623D, D627A, P628A, V630I, 1919A-3CGC, A644D, S649C, G653A, E654A, S661T, D669	P94L, M291I, V301L, H302R, A423V, V457L

gyrA	T2I, T5A, P7L, S11A, I15V, I20V, E21G, Q22R, R45G, F64L, D67G, A71V, A71G, A74S, A74V, N83K, G88C, G88A, D89N, D89G, **A90G**, **D94V**, S95G, **L105R**, D111N, T135I, M141T, F152L, I153M, P154A, P154R, P163A, S168I, L174V, **N193S**, E206Q, D211E, E214D, M220R, G221W, R222W, V223I, G239R, G239D, G249C, R254C, V259I, T267I, Q277R, N282K, F283L, S286L, A288D, V291L, R292Q, D293N, A322E, V323L, V327L, A343T, M345V, M345I, L346V, A347S, G351A, L356M, L358M, H368Q, R376S, R382L, E386V, H389Q, A406T, S411A, R418Q, Q431E, M438I, R448H, Q449E, A456V, A456G, I462V, E466Q, I468V, G477E, D488A, D488E, H490R, D493H, R495C, R495H, A500E, V505I, T519I, A547T, L568V, R578Q, A579V, D583E, E586V, R592S, Q594H, P604R, E605K, R607H, P621L, A626V, K633R, T638I, D639E, D641A, D641E, V651I, L653V, V663A, A676V, S684L, P692S, S698L, F705L, N706S, L711M, N715S, G729R, A736V, E739Q, V742L, Y755C, A765T, K793N, A814T, G823A, N826D, A827T, A827D, D829E, G832V, Q835*, T836K, T836M	P8A, E21Q, T80A, **A90V**, **S91P**, **D94N**, **D94H**, **D94Y**, **D94A**, **D94G**, S95T, G239S, G247S, S250A, R252L, L296P, A384V, R442H, A463S, P472S, Q613E, G668D
iniA	K19N, D22N, V33G, I46V, N50D, V71A, V74L, S84I, S84R, L87R, L91R, A96P, V98I, D100G, V105M, 321C-1A, V110I, V110L, 403A-6TTCCCG, D138E, D143N, E154A, P163R, S164R, L166M, 519T+1A, G180R, L191M, D198A, 623A-1G, V218M, A221T, V227A, V230I, V231M, R242W, A249V, M259T, I261V, L267M, H271D, E279K, 840C-1A, 848A-2GT, L293V, S294G, R300C, R306P, G308R, S323F, R339Q, 1041C-1A, Q357*, Q363R, R381Q, R385C, G408A, W425*, S430Y, D435N, A444T, S448P, 1460T-3GCG, Y491C, G505R, G507R, 1538C-4GGTG, 1541T-1G, 1563C-17CGGATGGCATATAAAGA, Y525C, 1597T-9TGCTGCGGG, V544A, V548F, D549N, 1655C+1A, S559A, T582I, 1762T+1C, A591T, V598A, R608W, Q611R, G615R	Q26*, **H42R**, 281C-5CGCGG, V106A, 516G-1T, F286C, L372P, Q394E, H481Q, S501W, R522Q
rpoB	D3G, S21F, N24D, G28R, P30S, P30A, S34A, R39S, **P45S**, **P45A**, **P45L**, P45R, D53N, E66K, A69P, V77M, E86K, P89L, D92G, F93V, M97L, 289A+6TGTCGT, V113I, M121V, M121I, E132D, M153T, V168A, **V170L**, **V170F**, Q172R, V179A, H194Y, S195R, F208L, R219C, Q226R, E244K, E250G, S254L, D259N, D265A, D270E, K274N, P280L, A286V, R299C, Y308C, N311D, L314V, A334D, H343Q, G345C, P358L, V359A, D362H,	**L430P**, **D435Y**, **D435G**, **D435V**, **H445N**, **H445D**, **H445Y**, **H445R**, **H445L**, **S450W**, **S450L**, **L452P**, **I491F**, A692T, **L731P**, R827C, V970A, Q975H, I1106T

	L378R, M390T, E391G, T399I, **T400A**, T400I, Q409R, F424L, F424V, T427P, 1281C-9AGCCAGCTG, S428G, S428R, Q429P, **Q429H**, **L430R**, **S431G**, Q432K, 1294C-9AATTCATGG, **Q432P**, **Q432L**, 1296A+3TTC, 1301T-6GGACCA, **M434I**, 1302G-3GAC, **D435A**, S441T, **S441L**, 1326G-6TTGACC, L443W, L443F, 1333C-3ACA, **H445P**, H445Q, **K446Q**, K446R, L449M, S450*, A451V, A451G, P454L, **E460G**, V469L, I480V, I480T, E481A, T482N, **I491L**, **I491V**, I491T, S493T, V496M, V496A, R511L, D515Y, V534M, V534L, V534A, V534G, A538V, S540A, A544V, D545A, D545E, 1648G+3AGC, R552S, R552H, R552L, L554P, V562A, E563D, V581M, M587T, H593Y, D634G, E639Q, E639G, R661Q, M666T, S672Y, H674Y, H674P, H674Q, G675D, T676P, T676A, P682T, V695L, T702I, H723Y, H723D, L741F, **E761D**, P802L, E812G, R824L, E825G, R827H, R827L, P834T, P834L, H835P, H835R, G836S, A857T, R871H, I873F, **S874Y**, **S874F**, G890D, L893R, P899A, I910T, M920V, I925V, K944E, R952G, P954H, E956D, V970M, V970L, G981D, A998V, H1028R, A1037S, Q1056H, Y1073S, Q1080R, V1096M, V1117L, S1124A, V1129A, E1169A, A1172P	
rpoC	G13R, L14R, Q22R, D44E, E49Q, E49A, I51V, D57N, M92T, T137A, E142G, H145Y, Q165R, Q165H, A172G, R173Q, E185Q, G188A, R203C, R211H, T225N, T227I, A230V, K232T, R247H, Q262R, D279G, V299A, V313A, A316T, **Q329K**, **G332S**, **G332R**, **G332C**, L402F, L402H, S403A, V431M, **G433S**, **G433C**, P434T, P434A, **P434Q**, **P434L**, **P434R**, Q435P, Q435H, G442C, **K445R**, A448V, L449V, F452S, F452C, F452L, R459W, A466V, W484C, D485N, D485Y, I491T, A492P, R506Q, L507V, L516V, V517L, E518A, E518D, **G519S**, **G519R**, **G519D**, A521D, A521V, Q523K, Q523E, H525N, M541I, M559L, S561P, L566V, G597D, P601Q, V611F, D623G, D623E, V629M, R641W, A648D, M663I, T667M, L679P, L679R, P682L, N698K, A701V, I707V, V709L, K715T, Y722C, A734G, P739L, R741S, R741C, D747H, D747G, E750G, **E750D**, R752H, K758Q, N766S, D768Y, L774V, L789V, L789S, T812I, **T825A**, N826T, N826K, E830A, A856T, **I885V**, 2665C+3ACG, T893I, E894G, E903A, P906A, D907G, E917D, E932K, T958I, G980E, M1012L, E1033A, E1033V, **V1039A**, **V1039G**, **P1040T**, **P1040A**, **P1040L**, F1061L, G1072D, E1106K, Q1110H, S1115L, **Q1125H**,	A172V, D271G, **N416S**, P481T, **V483A**, **V483G**, **W484G**, **I491V**, **L516P**, G594E, P601L, A621T, **N698S**, **P1040S**, **P1040R**, A1044V, E1092D, **K1152Q**, **V1252L**

	V1135A, E1140G, R1163C, V1206G, D1218A, S1242N, E1250A, **V1252M**, I1264T, V1272A, T1284A, S1287P, S1287*, S1287L, E1289A, A1303V, G1311S, Y1312H	
rpsL	P2R, S17R, K18E, E70K, **K88T**, **K88M**	**K43R**, **K88R**
rpsE	Q4H, R18W, R33C, R33H, F64S, N74S, G80C, K85T, L107V, H114Y, Q117R, A141G, A142V, A144S, T171A, P181L, P191Q, S207N, *221W	K39T, V105A
rrs	G5T, T17G, C22T, G38A, G102A, 115A+1T, T140G, 159C+1T, C171T, C181G, C182G, C196T, T200A, A208T, G261A, C270T, C332T, A335C, G349A, G361T, C380A, G395A, G395C, C397T, G406T, T454C, C462T, A484T, T529G, A554T, G583T, 642C+1T, C662G, G685A, C699A, C699G, C699T, A703G, G704A, C708T, G725A, G737T, A740C, A753T, G754A, G754T, G762A, C774A, A807C, T822C, T829C, C845A, C845T, C850T, G851T, G883A, G887T, G888A, G888C, C897T, C897G, **C905A**, **A906G**, A908C, A908G, A908T, G909T, G922A, C924T, G935A, T953C, A970C, A1012G, G1016C, G1016T, C1021T, T1025C, T1025G, G1026A, G1068A, C1105G, G1108T, C1125T, A1128T, 1144G+1T, T1151C, A1161G, G1167A, G1176A, A1205G, T1206C, T1208G, T1216C, C1220G, G1234A, G1237A, T1239A, T1239C, C1241T, A1244G, A1278C, A1278G, A1278T, G1285A, C1300T, G1302A, G1302C, C1319A, C1319G, G1321A, C1346T, G1353T, G1366T, G1379A, C1382G, **C1402A**, **C1402T**, T1444C, A1449G, G1450A, G1460A, A1461G, A1462C, A1469G, G1484T, C1489T	C282T, G284C, G292A, T305A, T327C, C492T, **C513T**, **A514C**, **A514T**, **C517T**, C710T, A753C, G771A, G878A, A899G, C936T, A948T, T958A, C1050T, 1075G+1T, T1208A, C1257T, C1357T, **A1401G**, C1507T
Rv1482c -fabG1	**T-8G**, G-9A, **C-34T**, G-77A, C-118G, C-120T	**T-8C**, **T-8A**, **C-15T**, **G-17T**, G-47C
inhA	I21V, I95L, G141R, T162S, G183R, E219A	I21T, **S94A**, T162A, **I194T**, I228V
tlyA	L26W, G28S, V32L, D35G, G36R, A45T, D48N, T56I, W62L, 198A+2GC, H68R, A80T, R84G, L87V, G95E, L100V, T134I, D154A, S156A, V163A, L164S, P179S, E186G, G188E, P194L, G196R, G196E, R206Q, H221R, P231R, N236K, 732C-10ACGCAGACCG, T247I, A253P, R262C, V1M	751T+2TG
katG	V1A, 8T-2CG, Q4*, Q4H, T12A, A16T, A16V, S17N, 55C-1G, V23L, 89A-3CGG, G34A, Q36P, V47I, D63G, A66T, V68G, T77R, R78G, V83G, M84T, Q88E, W90R, 269C+1A,	P6S, **S315T**, **S315N**, **S315R**, R463L, V473L

	W90*, G99R, F102S, R104Q, 317G+1C, A109T, A109V, G120S, G121S, G123R, G124D, G124A, 374C-2CG, Q127P, P131A, P131S, W135*, N138H, N138D, S140N, L141S, L141F, D142G, W149R, W149*, Y155S, Y155C, L159F, L159P, A162E, A162V, D163N, G169S, T180K, **D189A**, **D189G**, **W191R**, **W191G**, E192A, D194N, E195K, W198*, L205R, R209C, S211N, D215E, P232A, P232S, M242V, A244G, T251K, R254H, R254L, M257I, D259Y, V260I, T275A, G279D, D282G, G285V, 867C-6TCGGGT, A290V, Q295E, Q295P, G299S, S302R, T308A, S315G, **S315I**, I317T, 957G-1A, V320L, N323S, T324P, T324I, T326M, I335V, Y339S, E340D, E342G, T354I, L378P, A379T, T380P, **T380I**, S383A, L384R, D387H, T394A, L398R, D406G, F408L, W412*, Y413H, Y413C, D419H, D419Y, D419G, P422H, L427F, P432T, V445I, S446N, D448A, V450I, 1351C-12GACGAGGTCGTG, E452Q, 1366C-1A, Q471R, V473F, T475I, A480S, S481L, R484H, K488E, G495C, D509N, P510A, D511N, R519H, E522K, E523K, E523D, Q525K, Q525*, Q525P, S527L, A532P, A532V, G534R, K537E, 1612C+1T, D542E, C549S, A551S, 1671C-1T, K557N, N562H, G570C, G570V, L587R, A591T, L598R, 1811A-7ACGGGTT, A606T, M609T, D612G, T625A, G630V, V633A, 1901A+1G, Y638H, A649T, L653Q, T667I, K681T, S692R, D695A, L696Q, S700P, R705G, D729G, V739M	
pncA	M1T, **L4S**, I5T, 15G-3ATC, 16T+1G, I6M, V7I, V7L, V7F, V7G, D8N, **D8G**, D8E, V9A, **Q10***, 29T-1G, **Q10R**, 35T+1C, **D12A**, **D12G**, F13I, F13C, F13L, C14G, G17S, G17D, S18P, V21I, **V21A**, **V21G**, G24V, 72G-1C, 79G-11CGCGGCGCCAC, L27P, A28T, I31S, L35P, 105C-1A, E37*, E37V, 117C+1G, Y41*, H43P, **V44A**, **V44G**, A46P, A46E, A46V, **T47P**, **T47A**, T47I, K48T, **D49N**, D49A, D49G, D49E, H51D, H51Y, **H51P**, **H51R**, H51Q, D53E, **P54A**, **P54S**, P54Q, P54L, 165A-3CCC, H57D, H57Y, **H57P**, **H57R**, H57Q, F58S, F58L, S59P, T61P, P62S, **P62L**, D63H, **D63A**, **D63G**, **Y64D**, 193A+1T, S65P, 194G+1T, S66P, S66L, 201C-12GACGAGGAATAG, **W68R**, **W68G**, **W68***, **W68L**, W68C, P69T, P69S, P69R, P69L, 210C+3GGT, 210C+1G, H71Y, **H71Q**, C72R, C72Y, T76I, P77L, **G78S**, 232C-1G, **G78C**, G78A, A79T, F81V, H82D, 250T+1G, **L85P**, **L85R**, T87M, 277C-5CGCCT, F94L, F94C, Y95*, **K96T**, **K96R**, **G97S**, **G97R**, **G97C**, Y99*, A102P, Y103H,	**Q10P**, C14R, Y34D, L35R, T76P, G97D, A102V, **Y103***, L120P, **V125G**, **G132A**, I133T, **V139M**, **Q141P**, **L151S**, 457T+1G, 518T+1C

	310T-4GTAC, S104G, S104R, G105D, 315G+1C, F106L, T114M, 347A-9GTGGCGTGC, 347A+1G, L116R, R121P, R121Q, R123G, **V125D**, E127*, V128F, D129N, V130M, 389A-3CAT, 389A-9CATCGACCT, 390C-15ACATCGACCTCATCG, 390C-1A, 392A+1C, 392A+2CC, V131G, 395C-9CGACCACAT, 396A-1C, I133S, A134D, A134G, A134V, **T135P**, D136N, D136G, H137R, **C138R**, 415C-3ACA, V139L, **V139A**, V139G, 417C-1A, Q141*, 423C-1T, T142A, A143V, **A146T**, R154G, 464A+1C, **V155A**, **V155G**, 467A-8GCACCCTG, 471C+1T, L159P, L159R, T160P, T160A, G162S, G162D, S164*, 496C-1G, T168P, T168N, T168S, T168I, A171P, A171E, A171V, L172P, L172R, E173G, 521T+1A, M175K, M175T, M175I, A178P, V180L, V180F, **L182S**, **L182W**, V183L	
pncA-Rv2044c	C-33T, G-30T, G-19A, A-12G, **T-11G**, -2C+1G	**T-11C**
Rv2172c-idsA2	G-12A, A-55C, G-76T, G-97A, G-136T, T-149G, G-160A, A-173C, C-186T, A-205C, G-208A, T-234C, G-244A, G-259A, C-260T, G-265A, T-284C, T-284A	**A-65G**, C-98T
eis-Rv2417c	G-109A, A-106G, A-106C, C-104T, A-67T, G-63C, G-21A, G-15C, **C-10G**, -7C-1G, C-6A	G-100A, **G-14A**, **G-12A**, **C-10T**
oxyR'-ahpC	-3G+1T, -35C-1A, T-42C, -47G+1T, **C-52A**, C-54T, **C-57T**, T-71G, G-74A, T-76A, T-77G, C-79T, C-79A, A-80G, A-83G, A-98C	**G-48A**, **C-52T**, **C-72T**, **C-81T**, G-88A
thyX-hsdS.1	-239T+4CTAC, C-225T, A-206G, G-200T, C-176T, G-170A, C-167A, G-166T, G-152A, -127A-1G, T-117G, G-116A, A-108G, T-98C, G-58A, T-43G, G-42A, T-41G, A-31G, C-21T, **C-9T**, **G-4T**, **G-4A**	G-23C, **G-16A**
thyA	P3S, Y4*, **T22P**, **T22A**, 69G-5CCGGT, Q32H, V50A, A56V, L60M, H75N, G76*, I79T, W80*, D81A, D81G, 264G-1C, P92L, Y94C, **Q97R**, D117G, R120C, I128S, W133*, V135F, P145L, G157S, R158W, L159R, P175Q, 531G-1T, H207R, I208M, H254D, A259V, P260T, V263L	T202A, P253A
alr	F4Y, N12H, G19R, G19S, S22L, L23M, T26I, S29F, A38V, G71S, H72Y, T75A, T75M, P122S, D139H, D139G, E140D, T155A, V156A, K157E, T160A, D186G, A187S, D205G, A212D, F215V, A217V, S235W, S238L, T247M, L260V, S261N, P262Q,	F4L, **L113R**, M343T

	D268G, G270E, M275I, V284L, I287V, A308G, P311L, D316E, V318M, R325P, R340L, L350P, A363T, I364V, E373G, Y388D, Y388C, R397G, R397L, T401I, E406V	
embC-embA	**C-8A**, **C-11T**, **C-15G**, **C-16A**, G-17A, -27T-1A, T-27C, -29C-1T, -30C-2CT, G-32C, -33C-1G, -35A-1C, -38C-1T, G-48C, C-59A	**C-8T**, **C-11A**, **C-12T**, **C-12A**, **C-16G**, **C-16T**, **G-43C**
embB	R7T, R14Q, I16L, G37S, V50A, 177A-3CAG, G62R, V67L, I72L, I72S, L74P, D78G, D78E, P93L, G100S, P103T, K107R, S119N, V131M, V135M, R147C, E149A, F161L, K165N, R182C, V186A, V188A, P195H, A196T, T208I, A221T, A228V, V230A, V231I, A232P, L239V, G246R, L253I, A259V, G263R, W273L, V283M, F285L, N296H, S297A, D300G, G305C, **M306L**, M306T, D311A, S317F, Y319D, **Y319S**, **Y319C**, F323L, **D328H**, **D328Y**, F330V, W332R, M340I, T341A, T341N, T341I, H342N, S344R, L348P, M350T, D354N, C361Y, C361S, L370R, P375S, A386E, A388T, A388V, N399T, N400S, L402V, E405D, **G406C**, A409P, S412P, S422P, P430L, A438T, G443S, Q445R, A451T, M462L, M462T, R468H, I489T, I489S, **Q497P**, **Q497H**, T498N, A505T, A510T, S538P, T546A, T546I, A547S, M557I, L558F, K561R, I563L, V566M, G569A, V602I, G603R, R620C, F628S, L632F, L638F, W640S, T642A, W646L, P655Q, N657D, S658N, S658R, G665R, V668A, F676S, A679T, A680T, A693T, G694S, A701T, A716P, P731L, G748E, P776L, V783I, T797A, T797M, K820T, S823R, G836R, A840P, Q853P, S856R, D869H, D870N, P907S, G908R, A913V, Q925H, R930H, A943V, A950V, E951Q, L971M, 2942C-3GCA, M1000R, **H1002R**, I1006M, A1007V, K1011T, F1012L, D1017N, A1020S, L1037I, H1047P, V1048I, M1049I, D1056E, R1059P, T1069P, A1083T, W1089R, G1097S	L74R, Q139H, G156C, S203L, T205A, M306L, **M306V**, **M306I**, **D354A**, E378A, **P397T**, **G406S**, **G406D**, **G406A**, **M423T**, **Q497K**, **Q497R**, V668I, **D1024N**, S1054P, T1082A
ubiA	A15V, V18F, P23A, L31P, A35S, **A38T**, **A38S**, **A38V**, V44L, V55L, V55G, V61A, V78I, V105M, A106S, P122Q, M128L, G141S, L158S, I170V, L172P, **S173A**, **K174T**, **K174R**, W175C, F176S, **F176L**, **M180V**, M180I, A181T, A181V, T227I, A228S, **V229A**, **G234V**, A237T, A237V, R240C, R240P, D265Y, A278V, V283L, A300V	E149D, **V188A**, L224F, **A249T**, G268D
ethA	M1R, M1I, L5F, I9S, V10G, G13E, G16A, H22P, L23R, **Q24***, C27Y, Y32*, A33G, I34T, 102G-1A, 111C-1T, 119G+1T, G42S, G42D, G42V, G43V, T44A, 131G-1T, W45G,	H101R, R261W, S266R, T314I, N345K, A381P, E433A

	L47M, 141C-1A, F48S, 149T-1A, Y50C, 151G-1A, P51S, P51H, P51L, S55A, D56A, D56G, S57Y, Y60*, 181T-1G, T61K, 192G-1A, F66S, W69R, T70S, 230T-1C, G78D, P80T, L82P, A90V, I94S, R99W, 309C-1T, I105L, W109*, A112G, N114T, V118G, H119D, L129V, C131*, L136R, C137R, Y143*, Y147C, Y147*, P149S, G153S, P160T, P160L, H163Y, 492C+2GG, Q165*, H166P, D174G, N177S, V179F, S183R, T186R, V188I, T189K, P192T, P192S, L194P, S197*, 598T-1G, K200*, K200M, V202F, T203P, L205P, Q206*, 621G+1C, S208L, Y211S, Y211C, I212N, 641G-2AC, D219E, I221V, I221M, A222V, K224*, 673G+1C, 674A+2GC, L225P, W228R, P230L, 701G+2CC, R239G, A248T, 753G+1C, S251R, 756G+2GC, 757A+1G, Q254P, 771T+1G, 775G-1C, R259H, 788A-3TCT, F264L, E274K, 826A-1C, Y276H, Y276S, Y276*, H281P, H281R, P284L, 852C+1G, H285P, Y286D, Y286S, 861G-1T, P288R, 870G+1T, D290E, 885C-1A, 897G-1C, L301P, I305T, 955G-1T, R319W, F320S, P334A, I339N, A341V, 1023T+2GC, T342K, T342M, **Q347***, L348F, 1048C-1A, G351V, 1055G-1C, T353M, 1060C-3CGT, T355A, T355I, I356T, 1081C-1T, D362N, A368V, 1105A-1G, K370M, M373T, L374F, G376C, G376D, P378L, N379D, N379S, 1137G-1T, N379K, 1164A+1T, A389D, A389V, S390F, W391*, L397P, 1194C+1A, 1195A+3CAC, S399P, S399*, E400D, V402I, C403R, C403W, Y408D, 1225T-10GTAATTCAAC, G413D, F414C, 1243C-1A, G423R, F431V, 1304G+1T, P447A, K448E, G450D, T453I, P454S, P454L, W455R, R456P, Q459*, 1406C+1G, R469P, G471R, 1466A-1C	
ethA-ethR	A-3G, G-6A, G-21T, C-26T, A-29G, T-52A, A-60G, A-61G, A-69G, C-70G	**T-65C**
gid	P6A, P6R, **A8V**, A10V, I11N, R15P, L16P, L18H, A19P, A19G, A19V, 58G+1A, R20P, R20Q, R21W, R21P, A23D, L26S, L26F, A27P, G28E, 84T-1C, G30R, G30D, G30V, E32D, R33P, G34W, G34E, G34V, L35M, 103G+1C, 104A+1G, L35R, 108C-1A, G37A, G37E, E40*, V41L, V41I, V41G, L44R, W45R, W45*, 137T-1C, R47W, R47Q, H48N, H48D, H48Y, H48P, H48R, H48Q, L49P, L50P, L50R, C52R, C52Y, C52F, 158G-1C, L58F, L58P, L59V, L59F, E60*, R61P, R64W, R64P, R64Q, **V65A**, **V65G**, V66G, D67A, D67G, D67E, G69S, G69D, S70R, S70N, S70I, G71R, G71*, G71E, A72S, G73R, G73A,	A10P, L16R, 103G-1C, G37R, 116C-1G, S70R, G71V, P75S, **L79S**, **A80P**, Q87E, L90V, E92D, 352G-1C, A119T, 387G-1C, A138E, **S149R**, A167P, S181L, Y195H

	L74S, L74F, 223G-1C, P75A, **P75R**, **P75L**, 225C+1A, G76C, G76V, **V77A**, **V77G**, P78Q, P78R, P78L, **L79W**, L79F, A82P, R83G, R83W, R83P, R83Q, P84S, P84L, D85H, D85Y, D85A, L86V, L86F, L86P, Q87*, V88A, L90P, L91V, L91P, E92Q, E92K, E92G, P93S, R96H, R96L, R97H, R97L, 294G+2GT, 297C+1T, L101F, R102G, R102*, E103Q, E103*, E103A, T106I, L108P, L108R, 327G-1C, V110A, V112G, E113*, I114T, V115G, G117R, G117V, 352G+1C, R118S, 353C-1G, R118L, A119D, A119V, E120K, 358C-1G, E120*, S122A, 368C-2AG, V124G, Q125*, Q125P, Q127*, G130A, 391T-1G, S131R, A133P, A133G, 400C+1A, A134E, A134G, A134V, V135A, V135G, 405C-1A, S136P, S136*, R137W, R137P, R137Q, A138V, **V139L**, A140V, A141E, L142W, T146K, W148*, 446C+1T, S149R, 451G-1C, 453C-1G, L152V, L152S, 456C-1A, R154W, R154P, R154Q, G157R, 472G-1C, L160V, L160P, L160H, K163*, K163N, G164S, G164A, G164D, A167D, E170K, E173*, E173A, 519C+1T, R176C, A180G, S181*, G182R, V184F, 554T+1C, V186F, R187G, C191Y, C191F, 573A-14CATGTCACCACCCT, G192R, Y195F, Y195*, R197C, P199A, P199H, A200E, 602G-1T, V202A, V203L, V203G, F204L, A205P, A205T, A205E, R206G, R207H, R213P, R217G, R217W, M218V, M218I, A219V, A224G	
PPE52-nuoA	C-16T, G-27A, C-41T, C-55T, C-63T, G-80A, G-100A, G-122C, G-131C, G-154A, A-157C, C-217T, -231C-1G, A-249T, G-259A, G-261C, G-293T, G-300C, C-321T, G-329T, C-345G, T-346G, G-363C	**G-314T**
ald	M1T, M1I, N12D, E13K, R15Q, A17T, 128C+1G, 132T-1A, 133A+2TC, F50C, G54S, A55V, L57P, D62A, A68T, D69E, A98V, S100L, 317C+1T, T112I, T122I, L130P, 433G+2GC, M150T, Q153*, 459A-1G, A183V, A187T, M190T, 569T+1G, D198N, R214P, G227R, G237W, L240P, L249V, G261S, G261R, G261V, D266G, G277S, 837A+5CCGAC, P280L, 877A+2CG, L294Q, 896C+2GA, T308R, T317M, 966C+2GA, G328S, 991G+1C, A338E, P362L, S368G, S368N, 1106T+1G	A55P, Q56H

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Jody Phelan
Principal Supervisor	Taane Clark
Thesis Title	A Bioinformatic analysis of *M. tuberculosis* and host genomic data

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?	**International Jornal of Mycobacteriology**		
When was the work published?	June 2015		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I obtained the raw sequence data from our collaborators and proceeded to perform all analyses described in this paper. This process started with assessing the quality of the raw sequence data QC. After removing low quality data, I performed the genome assembly and improvement. Next, I annotated the genome and compared the sequences of drug resistance-related genes of *Mtb* and their homologues in *M. aurum*. Lastly, I performed the comparison of the proteome to other mycobacteria and performed phylogenetic reconstruction. I wrote all scripts to analyse the data and create figures using R and perl. I wrote the first draft of the manuscript and incorporated comments from co-authors. I, then, submitted to the journal.

Student Signature: ____ **Date:** __________________

Supervisor Signature: ____ **Date:** __________________

Chapter 5

The draft genome of *Mycobacterium aurum*, a potential model organism for investigating drugs against *Mycobacterium tuberculosis* and *Mycobacterium leprae*

HOSTED BY

ELSEVIER

Available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/IJMYCO

The draft genome of *Mycobacterium aurum*, a potential model organism for investigating drugs against *Mycobacterium tuberculosis* and *Mycobacterium leprae*

Jody Phelan [a,*], *Arundhati Maitra* [b,1], *Ruth McNerney* [a,1], *Mridul Nair* [c], *Antima Gupta* [b], *Francesc Coll* [a], *Arnab Pain* [c,2], *Sanjib Bhakta* [b,2], *Taane G. Clark* [a,d,2]

[a] *Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom*

[b] *Mycobacteria Research Laboratory, Institute of Structural and Molecular Biology, Department of Biological Sciences, Birkbeck College, University of London, Malet Street, London WC1E 7HX, United Kingdom*

[c] *Biological and Environmental Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia*

[d] *Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom*

ARTICLE INFO

Article history:
Received 27 April 2015
Accepted 3 May 2015
Available online 4 June 2015

Keywords:
Mycobacteria
M. aurum
M. tuberculosis
M. leprae
Drug screening
Genome

ABSTRACT

Mycobacterium aurum (*M. aurum*) is an environmental mycobacteria that has previously been used in studies of anti-mycobacterial drugs due to its fast growth rate and low pathogenicity. The *M. aurum* genome has been sequenced and assembled into 46 contigs, with a total length of 6.02 Mb containing 5684 annotated protein-coding genes. A phylogenetic analysis using whole genome alignments positioned *M. aurum* close to Mycobacterium *vaccae* and Mycobacterium *vanbaalenii*, within a clade related to fast-growing mycobacteria. Large-scale genomic rearrangements were identified by comparing the *M. aurum* genome to those of Mycobacterium *tuberculosis* and Mycobacterium *leprae*. *M. aurum* orthologous genes implicated in resistance to anti-tuberculosis drugs in *M. tuberculosis* were observed. The sequence identity at the DNA level varied from 68.6% for *pncA* (pyrazinamide drug-related) to 96.2% for *rrs* (streptomycin, capreomycin). We observed two homologous genes encoding the catalase-peroxidase enzyme (*katG*) that is associated with resistance to isoniazid. Similarly, two *embB* homologues were identified in the *M. aurum* genome. In addition to describing for the first time the genome of *M. aurum*, this work provides a resource to aid the use of *M. aurum* in studies to develop improved drugs for the pathogenic mycobacteria *M. tuberculosis* and *M. leprae*.

[*] Corresponding author at: Pathogen Molecular Biology Department, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom. Tel.: +44 (0) 20 7636 8636.
E-mail address: jody.phelan@lshtm.ac.uk (J. Phela).

[1] Joint authors.

[2] Joint authors.

Peer review under responsibility of Asian African Society for Mycobacteriology.
http://dx.doi.org/10.1016/j.ijmyco.2015.05.001
2212-5531/© 2015 Asian African Society for Mycobacteriology. Production and hosting by Elsevier Ltd. All rights reserved.

Introduction

Mycobacterium aurum (*M. aurum*) is an acid-fast, gram-positive environmental bacteria typically found in damp conditions [1,2]. It is a fast-growing mycobacterium with an *in vitro* doubling time of 2–3 h that rarely causes infections in humans [2–6]. The *M. aurum* cell wall contains mycolic acids which are analogous to those found in *Mycobacterium tuberculosis* [7], and there are similarities between the antibiotic susceptibility profiles of the two organisms [8,9]. The fast growth rate and low pathogenicity of *M. aurum* have encouraged its use as a surrogate for the highly pathogenic *M. tuberculosis* in studies of anti-microbial activity of anti-tubercular drugs [6,10,11]. Unlike other fast-growing mycobacteria, such as *Mycobacterium smegmatis, M. aurum* has the ability to survive within macrophages [12,13] and has been used for high throughput intracellular drug screening, allowing assessment of the ability of compounds to permeate the cell membrane and their stability within the cell [14,15]. The emergence of strains of *M. tuberculosis* resistant to multiple first- and second-line drugs threatens efforts to control tuberculosis (TB) and has renewed interest in the search for new anti-tubercular agents [16]. Rapid-growing models for screening putative anti-tubercular compounds are needed to accelerate drug discovery studies. Similarly, surrogate bacteria are needed to enable studies on drugs that may improve treatment for infection with non-culturable *Mycobacterium leprae*. Knowledge of the bacterial genome could enhance understanding of the molecular basis for drug resistance, and to this end, the genome of *M. aurum* has been sequenced and annotated. The genome was placed in a mycobacterium phylogeny, and comparisons with *M. tuberculosis, M. leprae* and *M. smegmatis* genomes were made in relation to susceptibility towards anti-tubercular drugs.

Materials and methods

M. aurum sample and DNA extraction

The *M. aurum* (NCTC 10437) was grown in 7H9 Middlebrook broth (Becton Dickinson, USA) supplemented with 10% albu min–dextrose–catalase (ADC) at 35 °C. DNA was extracted using the Bilthoven RFLP protocol [17]. In brief, log phase growth bacteria were treated with lysozyme, sodium dodecyl sulphate, proteinase K, N-cetyl-N,N,N-trimethyl ammonium bromide (CTAB) and chloroform-isoamyl alcohol prior to precipitation with isopropanol. Minimum inhibitory concentration (MIC) values for ethambutol, isoniazid, pyrazinamide and rifampicin drugs for the same *M. aurum* strain are available [18]. Duplications in *M. aurum* of *embB* and *katG* loci were confirmed by Sanger sequencing. For details of primers used, see Supplementary Table 1.

DNA sequencing and genome assembly

The *M. aurum* genomic DNA was sequenced using a 101 bp paired-end library on the Illumina HiSeq2000 platform. The raw sequence data (size 0.55 Gb, ~5.5 million paired reads, available from ENA ERP009288, minimum base call accuracy greater than 99%) underwent *de novo* assembly using *SPAdes* software [19]. The *SSPACE* software [20] was applied to scaffold the assembly, and a combination of *IMAGE* [21] and *GapFiller* [22] routines were used to further close or reduce the length of remaining gaps. An alternative approach using *Velvet* assembly software [23] led to a near identical assembly. Genomic annotation was transferred to the draft genome using the *Prokka* pipeline [24]. The pipeline searches for genes present in contigs and compares them with protein and DNA databases to annotate them. The *cd-hit* software [25,26] was used to integrate the annotation from 8 mycobacterial species to create a non-redundant blast "primary" database used by the *Prokka* pipeline. To validate the draft assembly and annotation pipeline, the transferred annotation was compared against the *kas* operon sequence (GenBank: DQ268649.2). All 5 genes from the GenBank entry (*fabD, acpM, kasA, kasB, accD6*) were annotated in the correct order and orientation in the assembly.

Comparative genomics

Genomes from 27 species used in whole genome comparisons were downloaded from *ensembl* (bacteria.ensembl.org), and the Uniprot taxon identification numbers are listed in Table 1. Gene multiple alignments were constructed using *clustalw2* [27] for 16S rRNA and *MACSE* software [28] for *rpoB* sequences. *Raxml* software [29] was used to construct the best scoring maximum likelihood tree, which was rooted using the *Corynebacterium glutamicum* (strain: ATCC 13032) reference sequence, an organism closely related to the mycobacterium genus [30]. Pairwise gene alignments were constructed using *MACSE* software, which uses the translated amino acid sequence and accounts for frame shifts and premature stop codons. Sequence identities were calculated using the *SIAS* webserver. Gaps were not used in the calculation of the percent identity. Whole genome alignments were constructed using *mercator* and *mavid* programs [31], and the resulting homology map was inspected and drawn using CIRCOS [32]. Orthologue clusters were created using *OrthoMCL* [33]. To identify any protein coding genes under selective pressure across *M. aurum*, *M. tuberculosis, Mycobacterium bovis – BCG, M. smegmatis,* and *M. leprae*, the *Ka/Ks* ratio was calculated, where *Ka* is the number of non-synonymous substitutions per non-synonymous site, and *Ks* is the number of synonymous substitutions per synonymous site. Ratio values less than one imply stabilizing or purifying selection, whilst values greater than one imply positive selection. To measure the degree of polymorphism across the genes, the nucleotide diversity (π) was also calculated using the same mycobacterial sample alignments. The *Ka/Ks* and π metrics were calculated using *variscan* (http://www.ub.edu/softevol/variscan) and *PAML* (http://abacus.gene.ucl.ac.uk/software/paml.html) software, respectively.

Table 1 – Genomic characteristics of *M. aurum* in the context of related species.

Organism	Chromosome accession number	Uniprot Strain taxon id	Assembled genome (bp)	G + C content	No. genes	Relative *in vitro* growth rate	ACDP risk class[a]
M. leprae	AL450380.1	272631	3268203	57.80	1605	Unculturable	3
C. glutamicum	HE802067.1	1204414	3309401	53.81	3099	Rapid	1
M. bovis	BX248333	233413	4345492	65.63	3952	Slow	3
M. tuberculosis	AL123456.3	83332	4411532	65.61	4047	Slow	3
M. xenopi	AJFI01000000	1150591	4434836	66.11	4281	Slow	2
M. canetti	HE572590.1	1048245	4482059	65.62	3981	Slow	3
M. thermoresistibile	AGVE01000000	1078020	4870742	69.02	4614	Rapid	1
M. hassiacum	AMRA01000000	1122247	5000164	69.46	4959	Rapid	1
M. abscessus	CU458896.1	36809	5067172	64.15	4942	Rapid	1
M. inracellulare	CP003322.1	487521	5402402	68.10	5144	Slow	2
M. neoaurum	CP006936.1	700508	5438192	66.88	4217	Rapid	1
M. avium	CP000479.1	243243	5475491	68.99	5120	Slow	2
M. gilvum	CP002385.1	278137	5547747	67.86	5349	Rapid	1
M. colombiense	AFVW02000000	1041522	5579559	68.09	5197	Slow	1
M. indicus pranii	CP002275.1	1232724	5589007	68.03	5254	Rapid	1
M. ulcerans	CP000325.1	362242	5631606	65.47	4160	Slow	3
M. yongonense	CP003347.1	1138871	5662088	67.90	5390	Slow	1
M. phlei	AJFJ01000000	1150599	5681954	69.21	5435	Rapid	1
M. aurum	**TBA**[b]	**TBA**	**6019822**	**67.52**	**5684**	**Rapid**	**1**
M. vaccae	ALQA01000000	1194972	6245372	68.60	5949	Rapid	1
M. chubuense	CP003053.1	710421	6342624	68.29	5843	Rapid	1
M. fortuitum	ALQB01000000	1214102	6349738	66.21	6241	Rapid	2
M. vanbaalenii	CP000511.1	350058	6491865	67.79	5979	Rapid	1
M. parascrofulaceum	ADNV01000000	525368	6564171	68.45	6456	Slow	1
M. kansasii	CP006835.1	557599	6577228	66.23	5866	Slow	2
M. marinum	CP000854.1	216594	6636827	65.73	5452	Slow	2
M. smegmatis	CP000480.1	246196	6988209	67.40	6938	Rapid	1
M. rhodesiae	AGIQ01000000	931627	7281599	66.07	7024	Rapid	1

a UK Advisory Committee on Dangerous Pathogens (ACDP) http://www.hse.gov.uk/pubns/misc208.pdf.

b ENA number ERP009288.

Results

The *M. aurum* genome

A total of ~5.5 million high quality paired end (101 bp) reads were used to assemble the *M. aurum* genome. The final *M. aurum* assembly consisted of 46 contigs, 43 of which were over 500 bp in length. The median contig length (N50) was 265 Kb (minimum 315 bp, maximum 742,983 bp). The total genome length (~6.02 Mb, G + C content 67.52%) is longer than *M. tuberculosis* (4.4 Mbp) and *Mycobacterium canetti* (4.5 Mbp), but shorter than *Mycobacterium marinum* (6.6 Mbp) and *M. smegmatis* (7.0 Mbp) (Table 1). A total of 5684 coding sequences, 1 tmRNA, 4 rRNA and 51 tRNA features were annotated, and of these 4306 (75%) were assigned a function (Fig. 1). The final contigs and annotation are available for download (pathogenseq.lshtm.ac.uk/m_aurum).

M. aurum and the mycobacteria phylogeny

A phylogenetic analysis using 27 mycobacterial whole genome sequences revealed that *M. aurum* clustered with *Mycobacterium vaccae* and *Mycobacterium vanbaalenii* within a clade related to fast-growing mycobacteria (Fig. 2). Slow-growing bacteria, including *M. tuberculosis*, clustered within a distinct clade. However, *Mycobacterium indicus pranii*, a fast-growing mycobacterium and immunotherapy and vaccine candidate for leprosy and tuberculosis [34], clustered within the slow-growing clade. The very high bootstrap support values for the phylogenetic tree (median 100%, range 77–100%) indicates the high precision afforded when using whole genome data. Previously, *hsp65, sodA, recA, rpoB* and 16S rRNA gene sequence data were used to barcode bacteria, with the latter approach being adopted widely [35]. The assembled 16S rRNA sequence for *M. aurum* had the highest identity with *M. vanbaalenii* (99%), *Mycobacterium rhodesiae* (99%), and *Mycobacterium austroafricanum* (99%), in concordance with previous reports [36,37]. The phylogenetic tree constructed using 16S rRNA sequences was broadly similar to that from whole genome data (Supplementary Fig. 1). However, *M. aurum* and *M. vanbaalenii* clustered closer to *Mycobacterium abscessus* rather than *Mycobacterium gilvum*, and the topology was less robust with lower bootstrap support values.

Comparison to the M. tuberculosis and M. leprae genomes

The *M. aurum* assembled contigs were ordered according to the *M. tuberculosis* H37Rv reference genome (AL123456.3), leading to 10 gapped scaffolds. Most of the *M. tuberculosis*

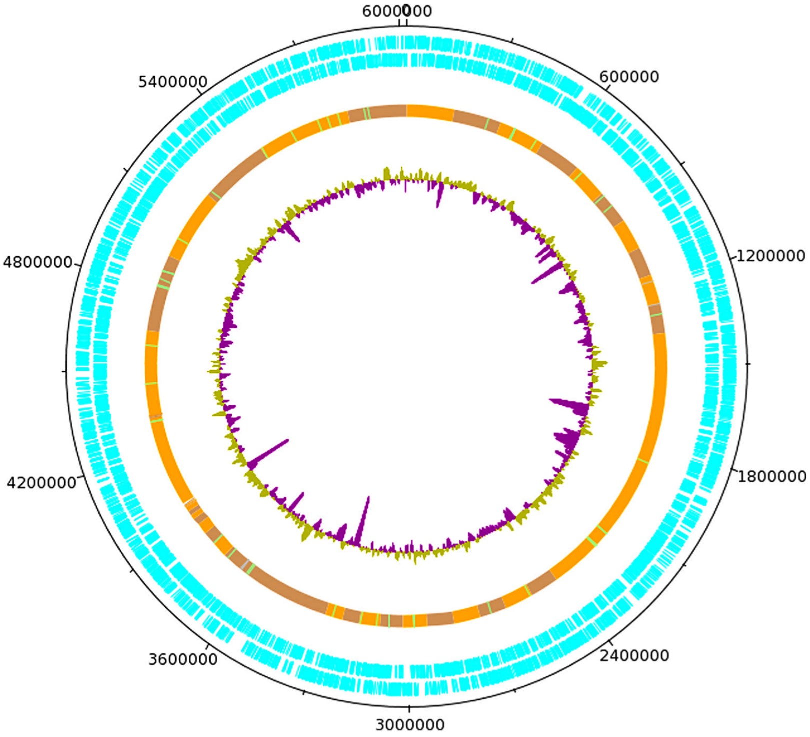


Fig. 1 – An annotated circular view of the *M. aurum* genome (length ~6.02 Mb). Innermost track: G + C% content; middle track: the 46 contigs, alternating between brown and orange with green and grey lines representing tRNA and rRNA, respectively; outer track: the 5684 forward and reverse genes.

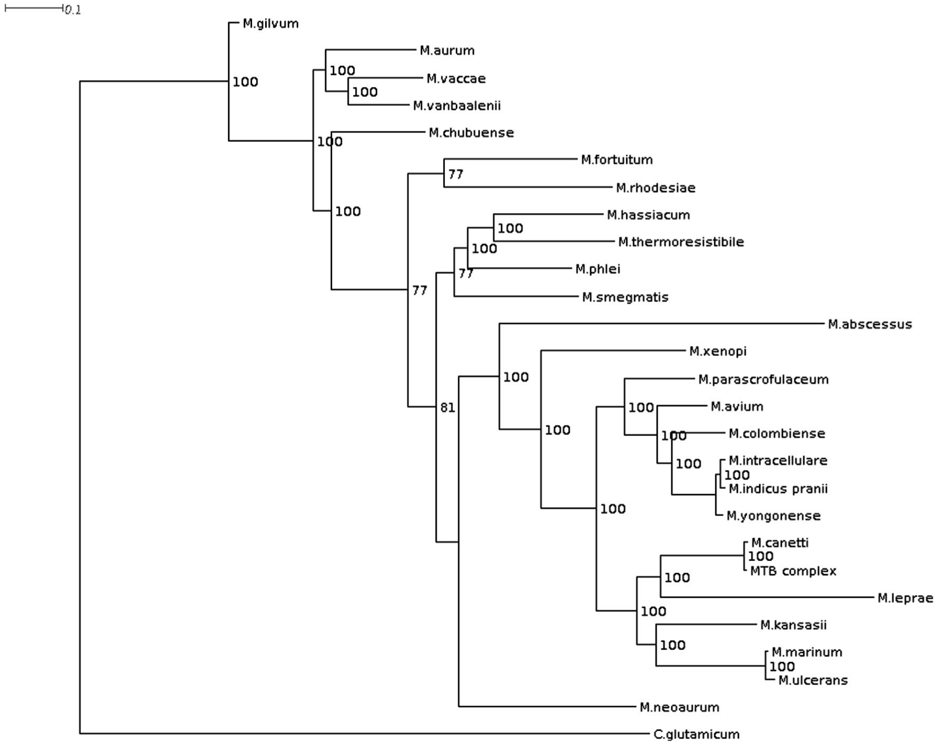


Fig. 2 – *M. aurum* and the mycobacterium phylogeny* constructed using 27 whole genome reference sequences. *Constructed using *RaXML* and statistic support for lineages was based on 100 bootstrap samples. 27 reference sequences used are described in Table 1.

genome (86%) had regions with synteny in *M. aurum*. The map of homology between the 10 *M. aurum* scaffolds and the *M. tuberculosis* genome consisted of 67 regions of synteny (Supplementary Table 2 and Fig. 3a). Although there was high similarity between *M. aurum* and *M. tuberculosis*, there was evidence for large-scale rearrangements (Fig. 3a). Twenty-eight genes required for survival within macrophages were observed, but a further two (*lpqY* and *eccA*$_1$) could not be found [38] (Supplementary Table 3). The putative proteome for *M. aurum* suggests it lacks 1002 proteins present in *M. tuberculosis*, but has an additional 2090 proteins not seen in *M. tuberculosis* (see Table 2).

The map of homology between the 10 *M. aurum* scaffolds and the *M. leprae* genome consisted of 73 segments of synteny (Supplementary Table 2 and Fig. 3b). For *M. aurum* and *M. leprae* there were 2047 and 222 unique proteins, respectively, which had no orthologue in the other mycobacteria (see Supplementary Table 2). *M. smegmatis* is often used as a fast-growing model of *M. tuberculosis*. A similar analysis carried out between *M. tuberculosis* and *M. smegmatis* revealed 979 and 2314 unique proteins for each, respectively, which had no orthologue in the other mycobacteria. When compared with the *M. aurum*–*M. tuberculosis* analysis, the number of apparently unique proteins in *M. smegmatis* was higher by 224 proteins.

Drug resistance candidate genes

Pairwise alignments were constructed for the known drug target genes to establish the degree of homology between *M. aurum* and *M. tuberculosis* (Table 2). The sequence identity at the DNA level varied from 68.6% for *pncA* (pyrazinamide drug-related) to 96.2% for *rrs* (streptomycin, capreomycin). The percentage of amino acid identity was higher than the sequence identity, being high among all drug resistance candidate genes analysed (range 90.6–99.2%). Interestingly, two genes at different locations were annotated as *katG* in the *M. aurum* genome, and denoted as *katG*1 and *katG*2. The percent identity between the two genes and their *M. tuberculosis* homologue at the DNA level are 73.6% and 68.8% (Supplementary Fig. 2) The putative *M. aurum katG*1 found in contig 20 (*aurum*03417) demonstrated the highest homology to the *M. tuberculosis katG* gene (Rv1908c) and *M. smegmatis MSMEG_6384*. The second *M. aurum katG*2 *(aurum* 02416) located in contig 2 (*katG*2) was most homologous with *M. smegmatis MSMEG_3461*. A third *M. smegmatis* gene, *MSMEG_3729*, showed weak homology to each of the *katG* genes in *M. aurum* and *M. tuberculosis*. Two copies of *embB*, a gene associated with ethambutol in *M. tuberculosis*, were also found in different locations in *M. aurum* (72.3% and 47.7% identity). The semi-identical duplications for each of *katG*

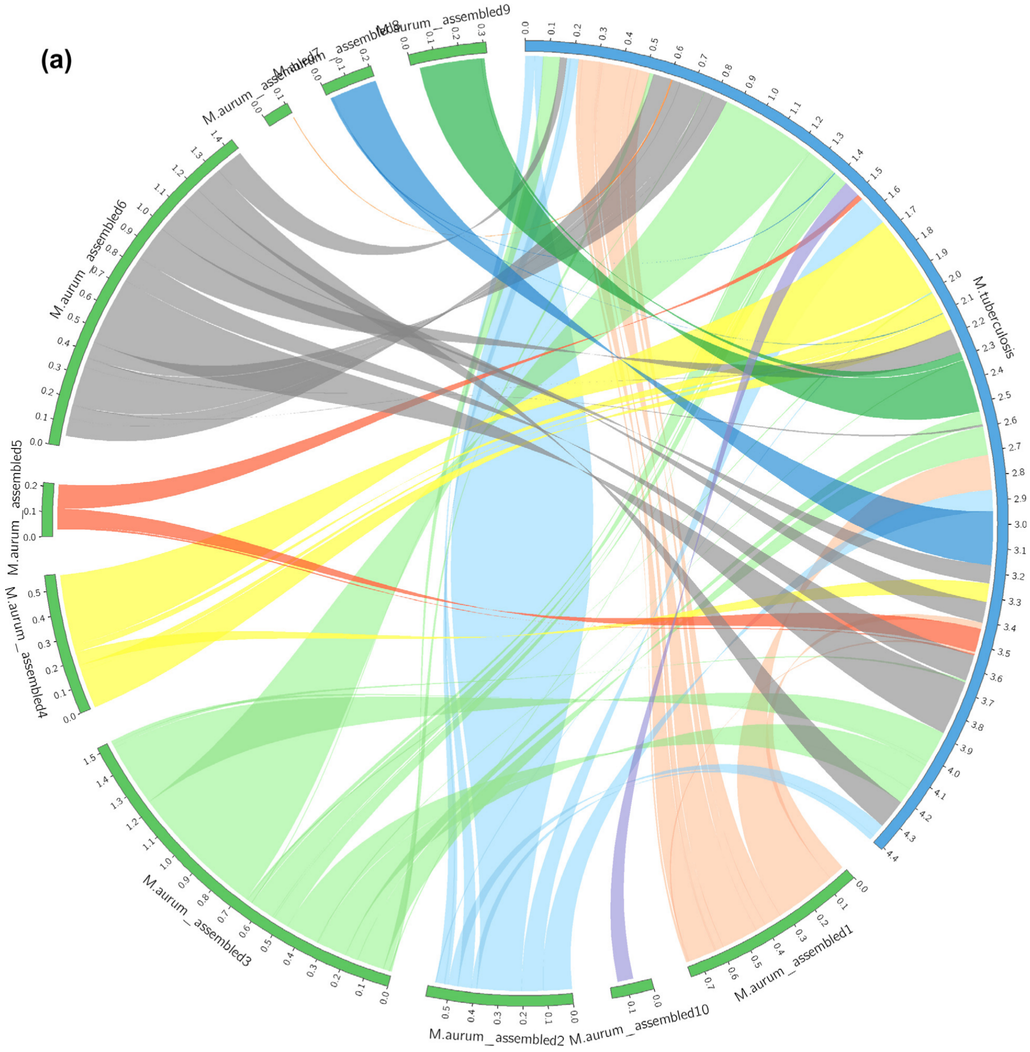


Fig. 3 – Homology between *M. aurum* and *M. tuberculosis* and *M. leprae*. (a) *M. aurum* (green) and *M. tuberculosis* H37Rv (blue). The ten contigs provide 67 segments of synteny with *M. tuberculosis* H37Rv. The segments range from 2,266 bp to 391,674 bp in length. (b) *M. aurum* (green) *and M. leprae* (blue). The ten contigs provide 73 segments of synteny with *M. leprae*. The segments range from 2,495 bp to 193,922 bp in length.

and *embB* were confirmed by PCR and Sanger sequencing (Supplementary Table 4).

Across a range of therapeutic agents, potential differences in minimum inhibitory concentration (MIC) levels between *M. tuberculosis* (H37Rv) and *M. aurum* for isoniazid, ethambutol and ofloxacin (Table 2) are available [8,18], with the biggest difference for isoniazid. The MIC values for isoniazid were greatest in *M. smegmatis* (2 mg/L), followed by *M. aurum* (0.4) and *M. tuberculosis* (0.02–0.2). No known *M. tuberculosis* mutations were identified in the *katG, inhA* (isoniazid), *ethA, ethR* (ethambutol), and *gyrA/B* (ofloxacin) orthologues in *M. aurum*. Homologues of *ahpC* and *embR* genes, associated with isoniazid and ethambutol drug resistance respectively, were not observed in the *M. aurum* genome.

The alignments were compared across *M. aurum*, *M. tuberculosis*, *M. bovis – BCG*, *M. smegmatis*, and *M. leprae* at the loci considered drug targets or those loci considered to have important functional roles (Table 3). All loci had a high percentage (~90%) of their nucleotides analyzable across the mycobacteria, except *fas* and *gyrA* where there were large insertions in *M. aurum* and *M. leprae*, respectively. Only three loci did not have alignment gaps: *inhA* (isoniazid drug-

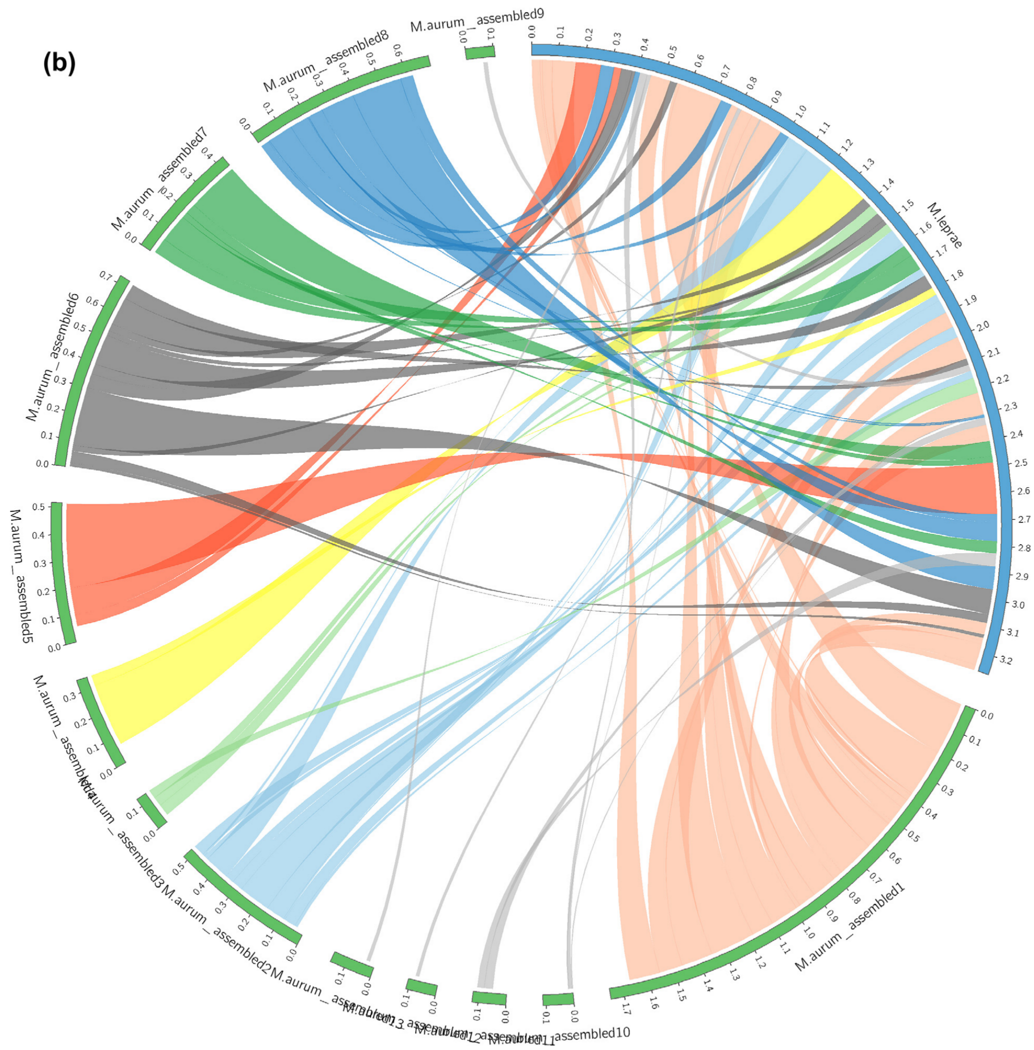


Fig 3. *(continued)*

related); *rpsL* (streptomycin); and *kasA* (thiolactomycin). The *ddn* (delamanid), *fpol1* (para-aminosalicylic acid), *murC/D/E/F* family (isoquinolines), and *nat* (cholesterol metabolism) loci were the most polymorphic (>40% sites segregating, nucleotide diversity $\pi > 0.2$). In contrast, the *rrs* gene associated with streptomycin drug resistance was the most conserved (2.9% segregating sites, pairwise diversity $\pi = 0.029$). In general, there was a modest degree of conservation in most genes (all with >50% of sequence conserved), which would be expected given the known synergistic drug effects across mycobacteria. All candidate genes reported *Ka/Ks* values much lower than 1, consistent with the selective removal of alleles that are deleterious (purifying selection). The highest *Ka/Ks* value was observed for *nat* (*Rv3566c*), a gene encoding arylamine acetylase that is associated with resistance to isoniazid [39].

Discussion

The draft genome sequence of *M. aurum* (length ~6.02 Mb, G + C content 67.52%) has been assembled. The genome assembly consists of 46 contigs and provides the first insight into the genetic code of *M. aurum*. Lack of alternative sequence data for this bacterium, particularly from technologies with longer reads, prevents closure of the gaps at this time. Using whole genome alignments, the placement of *M. aurum* within the mycobacterial phylogeny, close to *M. vaccae* and *M. vanbaalenii*, was confirmed. The analysis of loci

Table 2 – Drug minimum inhibitory concentrations (MICs) and candidate resistance gene identity between *M. aurum* and *M. tuberculosis* at drug resistance loci.

Drug	MIC[b] *M. aurum* mg/L (μM)	MIC[b] *H37Rv* mg/L (μM)	*M. tb* loci	Gene homology with *M. aurum* (%)	Protein similarity score (%)	*M. aurum* feature
Isoniazid			*katG*[a]	72.06	93.23	2 loci
	0.40	0.02–0.2	*inhA*[a]	87.40	98.14	
	(3.65)	(0.15–1.46)	*ahpC*	–	–	Absent
			kasA	86.33	99.04	
Rifampicin	0.10	0.10	*rpoB*[a]	90.74	97.62	
	(0.12)	(0.12)	*rpoC*	90.20	98.10	
Ethambutol			*embB*	69.87	91.46	2 loci
	0.5	0.47	*embA*	68.87	94.28	
	(2.45)	(2.30)	*embC*	73.83	93.78	
			embR	–	–	Absent
Streptomycin,	0.2 (0.34)	0.1–0.5 (0.17–0.86)	*rrs*	–	–	
aminoglycosides,	–	–	*rpsL*	96.00	99.20	
capreomycin	–	–	*tlyA*	73.58	92.98	
Pyrazinamide	>100,	>100	*pncA*	64.17	90.62	
	(812.26)	(812.26)	*rpsA*	93.56	98.55	
Ethionamide	5	0.6–2.5	*ethA*	65.71	94.30	
	(30.08)	(3.6–15.04)	*ethR*	69.15	93.11	
Ofloxacin	0.2	1–2	*gyrA*	90.34	97.75	
	(0.55)	(2.77–5.53)	*gyrB*	86.39	92.30	

Homology as calculated using protein alignment. Protein similarity is quite high for most proteins analysed.

[a] Selected alignments can be found in http://pathogenseq.lshtm.ac.uk/m_aurum/.

[b] MIC value Ref. [18].

Table 3 – A comparison across *M. aurum*, *M. tuberculosis*, *M. bovis – BCG*, *M. smegmatis*, and *M. leprae* alignments at drug targets or other important loci.

Drug resistance or function	Gene name	Alignment length[a]	% Sites analysed[b]	Gaps	% Segregating sites	% Conserved sites	π[c]	Ka/Ks[d]
Bedaquiline (TMC207)	*atpE*	261	94.25	15	28.0	72.0	0.150	0.089
BTZ043, DNB1, VI-9376, 377790, TCA1	*dprE1*	1410	98.09	27	34.6	65.4	0.210	0.128
Cholesterol metabolism	*hsaA*	1191	99.50	6	27.6	72.4	0.166	0.150
	hsaB	570	98.95	6	26.2	73.8	0.157	0.182
	hsaC	903	99.67	3	30.1	69.9	0.182	0.104
	hsaD	921	93.81	57	30.9	69.1	0.189	0.129
	nat	861	95.82	36	43.8	56.2	0.268	0.287
Fluoro-quinolones	*gyrA*	3807	65.33	1320	32.2	67.8	0.171	0.070
	gyrB	2157	93.88	132	36.2	63.8	0.193	0.077
Isoniazid/pyridomycin	*inhA*	810	100	0	28.1	71.9	0.152	0.101
Isoquinolines	*murC*	1512	94.64	81	43.1	56.9	0.235	0.170
	murD	1509	96.02	60	46.1	53.9	0.254	0.240
	murE	1653	91.83	135	48.1	51.9	0.267	0.228
	murF	1617	91.65	135	42.7	57.3	0.230	0.167
Isoxyl (thiocarlide)	*fas*	10701	85.79	1521	37.0	63.0	0.195	0.163
PA-824, delamanid (OPC67683)	*ddn*	513	88.30	60	44.6	55.4	0.280	0.331
para-aminosalicylic acid	*folP1*	882	92.18	69	42.9	57.1	0.241	0.144
	folP2	957	88.71	108	35.8	64.2	0.198	0.129
Q203, IP3	*qcrB*	1695	96.46	60	31.5	68.5	0.174	0.126
Rifampicin	*rpoB*	3537	98.47	54	23.7	76.3	0.128	0.084
Streptomycin	*rpsL*	375	100	0	23.2	76.8	0.122	0.028
	rrs	1563	95.84	65	2.9	97.1	0.029	0.029
Thiolactomycin	*kasA*	1251	100	0	29.1	70.9	0.159	0.100
	kasB	1326	91.63	111	36.4	63.6	0.195	0.117

Selected alignments can be found at pathogenseq.lshtm.ac.uk/m_aurum/.

[a] The total number of columns in the alignment including gaps.

[b] A function of the number of sites used in determining the number of segregating and conserved sites.

[c] π nucleotide diversity.

[d] The *Ka/Ks* is the ratio of the number of non-synonymous substitutions per non-synonymous site (*Ka*) to the number of synonymous substitutions per synonymous site (*Ks*).

involved in drug resistance demonstrated homology with *M. tuberculosis* and *M. leprae*. This insight corroborates earlier investigations of *inhA* gene mutants of *M. aurum* that showed similarity in drug resistance mechanisms against isoniazid and ethionamide between *M. aurum* and *M. tuberculosis* [6,40]. The draft *M. aurum* genome is larger than that of *M. tuberculosis* with an additional 2090 genes not observed in *M. tuberculosis*; it is also lacking 1002 of the genes found in *M. tuberculosis*. Multiple copies of some homologous genes were observed. Of particular interest are two putative copies of *embB*, a gene involved in the biosynthesis of the mycobacterial cell wall component arabinan and that is associated with resistance to ethambutol in *M. tuberculosis*. Similarly, two annotated catalase-peroxidase (*katG*) genes that may be involved in the activation of the anti-tuberculosis pro-drug isoniazid were identified and confirmed. Multiple *katG* genes have been reported in other mycobacteria, for example in *Mycobacterium fortuitum* [41]. It could be hypothesized that the duplications of *katG* in *M. aurum* and *M. smegmatis* could have an effect on the MIC values. Further laboratory work is underway to elucidate the endogenous function of the observed duplications.

In summary, these genomic analyses support the use of *M. aurum* as a potential model organism for providing insights into *M. tuberculosis* biology, particularly for new drug development, with the possibility of leading to new control measures for tuberculosis disease. Further insight may be gained from the genome sequence of additional strains and related mycobacteria.

Conflicts of interest

The authors declare no conflict of interests.

Acknowledgements

T.G.C. is funded by the Medical Research Council UK (Grant No. MR/K000551/1). J.P. is supported by a BBSRC UK Ph.D. studentship. The project was supported by the KAUST faculty baseline research fund to A.P.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ijmyco.2015.05.001.

REFERENCES

[1] S. Hartmans, J.A.M. de Bont, E. Stackebrandt, The genus mycobacterium – nonmedical, in: M. Dworkin, S. Falkow (Eds.), The Prokaryotes: Vol. 3: Archaea. Bacteria: Firmicutes, Actinomycetes, Springer Science & Business Media, New York, 2006, pp. 889–918.

[2] B. Honarvar, H. Movahedan, M. Mahmoodi, F.M. Sheikholeslami, P. Farnia, *Mycobacterium aurum* keratitis: an unusual etiology of a sight-threatening infection, Braz. J. Infect. Dis. 16 (2012) 204–208.

[3] J. Esteban, R. Fernandez-Roblas, A. Roman, A. Molleja, M.S. Jimenez, F. Soriano, Catheter-related bacteremia due to *Mycobacterium aurum* in an immunocompromised host, Clin. Infect. Dis. 26 (1998) 496–497.

[4] K.I. Koranyi, M.A. Ranalli, *Mycobacterium aurum* bacteremia in an immunocompromised child, Pediatr. Infect. Dis. J. 22 (2003) 1108–1109.

[5] A. Martin-Aspas, F. Guerrero-Sanchez, P. Garcia-Martos, E. Gonzalez-Moya, F. Medina-Varo, J.A. Giron Gonzalez, Bilateral pneumonia by *Mycobacterium aurum* in a patient receiving infliximab therapy, J. Infect. 57 (2008) 167–169.

[6] A. Gupta, S. Bhakta, S. Kundu, M. Gupta, B.S. Srivastava, R. Srivastava, Fast-growing, non-infectious and intracellularly surviving drug-resistant *Mycobacterium aurum*: a model for high-throughput antituberculosis drug screening, J. Antimicrob. Chemother. 64 (2009) 774–781.

[7] J.T. Belisle, V.D. Vissa, T. Sievert, K. Takayama, P.J. Brennan, G.S. Besra, Role of the major antigen of *Mycobacterium tuberculosis* in cell wall biogenesis, Science 276 (1997) 1420–1422.

[8] B. Phetsuksiri, A.R. Baulard, A.M. Cooper, D.E. Minnikin, J.D. Douglas, G.S. Besra, et al, Antimycobacterial activities of isoxyl and new derivatives through the inhibition of mycolic acid synthesis, Antimicrob. Agents Chemother. 43 (1999) 1042–1051.

[9] F.G. Winder, Mode of action of the antimycobacterial agents and associated aspects of the molecular biology of mycobacteria, in: C. Ratledge, J. Standford (Eds.), The Biology of Mycobacteria, Academic Press Inc., New York, 1982, pp. 353–438.

[10] R. Srivastava, D. Kumar, B.S. Srivastava, Recombinant *Mycobacterium aurum* expressing *Escherichia coli* beta-galactosidase in high throughput screening of antituberculosis drugs, Biochem. Biophys. Res. Commun. 240 (1997) 536–539.

[11] G.A. Chung, Z. Aktar, S. Jackson, K. Duncan, High-throughput screen for detecting antimycobacterial agents, Antimicrob. Agents Chemother. 39 (1995) 2235–2238.

[12] A. Gupta, A. Kaul, A.G. Tsolaki, U. Kishore, S. Bhakta, *Mycobacterium tuberculosis*: immune evasion, latency and reactivation, Immunobiology 217 (2012) 363–374.

[13] R. Srivastava, D.K. Deb, K.K. Srivastava, C. Locht, B.S. Srivastava, Green fluorescent protein as a reporter in rapid screening of antituberculosis compounds in vitro and in macrophages, Biochem. Biophys. Res. Commun. 253 (1998) 431–436.

[14] A. Gupta, S. Bhakta, An integrated surrogate model for screening of drugs against *Mycobacterium tuberculosis*, J. Antimicrob. Chemother. 67 (2012) 1380–1391.

[15] D.K. Deb, K.K. Srivastava, R. Srivastava, B.S. Srivastava, Bioluminescent *Mycobacterium aurum* expressing firefly luciferase for rapid and high throughput screening of antimycobacterial drugs in vitro and in infected macrophages, Biochem. Biophys. Res. Commun. 279 (2000) 457–461.

[16] A. Zumla, I. Abubakar, M. Raviglione, M. Hoelscher, L. Ditiu, T.D. McHugh, et al, Drug-resistant tuberculosis-current dilemmas, unanswered questions, challenges, and priority needs, J. Infect. Dis. 205 (Suppl. 2) (2012) S228–240.

[17] J.D. van Embden, M.D. Cave, J.T. Crawford, J.W. Dale, K.D. Eisenach, B. Gicquel, et al, Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology, J. Clin. Microbiol. 31 (1993) 406–409.

[18] J.D. Guzman, D. Evangelopoulos, A. Gupta, K. Birchall, S. Mwaigwisya, B. Saxty, et al, Antitubercular specific activity of ibuprofen and the other 2-arylpropanoic acids using the HT-SPOTi whole-cell phenotypic assay, BMJ Open 3 (2013).

[19] A. Bankevich, S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, et al, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, J. Comput. Biol. 19 (2012) 455–477.

[20] M. Boetzer, C.V. Henkel, H.J. Jansen, D. Butler, W. Pirovano, Scaffolding pre-assembled contigs using SSPACE, Bioinformatics 27 (2011) 578–579.

[21] I.J. Tsai, T.D. Otto, M. Berriman, Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps, Genome Biol. 11 (2010) R41.

[22] M. Boetzer, W. Pirovano, Toward almost closed genomes with GapFiller, Genome Biol. 13 (2012) R56.

[23] D.R. Zerbino, E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs, Genome Res. 18 (2008) 821–829.

[24] T. Seemann, Prokka: rapid prokaryotic genome annotation, Bioinformatics 30 (2014) 2068–2069.

[25] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (2012) 3150–3152.

[26] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (2006) 1658–1659.

[27] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, et al, Clustal W and Clustal X version 2.0, Bioinformatics 23 (2007) 2947–2948.

[28] V. Ranwez, S. Harispe, F. Delsuc, E.J. Douzery, MACSE: Multiple Alignment of Coding Sequences accounting for frameshifts and stop codons, PLoS One 6 (2011) e22594.

[29] A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, Bioinformatics 30 (2014) 1312–1313.

[30] M.V. Omelchenko, Y.I. Wolf, E.K. Gaidamakova, V.Y. Matrosova, A. Vasilenko, M. Zhai, et al, Comparative genomics of *Thermus thermophilus* and *Deinococcus radiodurans*: divergent routes of adaptation to thermophily and radiation resistance, BMC Evol. Biol. 5 (2005) 57.

[31] C.N. Dewey, Aligning multiple whole genomes with Mercator and MAVID, Methods Mol. Biol. 395 (2007) 221–236.

[32] S.J. McKay, I.A. Vergara, J.E. Stajich, Using the Generic Synteny Browser (GBrowse_syn), Curr. Protoc. Bioinformatics (2010) (Chapter 9: Unit 9 12).

[33] S. Fischer, B.P. Brunk, F. Chen, X. Gao, O.S. Harb, J.B. Iodice, et al, Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups, Curr. Protoc. Bioinformatics (2011) (Chapter 6: Unit 6 12 11–19).

[34] S.A. Zaheer, R. Mukherjee, B. Ramkumar, R.S. Misra, A.K. Sharma, H.K. Kar, et al, Combined multidrug and *Mycobacterium* w vaccine therapy in patients with multibacillary leprosy, J. Infect. Dis. 167 (1993) 401–410.

[35] T. Adekambi, M. Drancourt, Dissection of phylogenetic relationships among 19 rapidly growing *Mycobacterium* species by 16S rRNA, hsp65, sodA, recA and rpoB gene sequencing, Int. J. Syst. Evol. Microbiol. 54 (2004) 2095–2105.

[36] N.C. Gey van Pittius, S.L. Sampson, H. Lee, Y. Kim, P.D. van Helden, R.M. Warren, Evolution and expansion of the *Mycobacterium tuberculosis* PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions, BMC Evol. Biol. 6 (2006) 95.

[37] M. Goodfellow, J.G. Magee, Taxonomy of mycobacteria, in: P.R.J. Gangadharam, P.A. Jenkins (Eds.), Mycobacteria: Basic Aspects, Chapman and Hall, New York, 1998, pp. 1–53.

[38] J. Rengarajan, B.R. Bloom, E.J. Rubin, Genome-wide requirements for *Mycobacterium tuberculosis* adaptation and survival in macrophages, Proc. Natl. Acad. Sci. USA 102 (2005) 8327–8332.

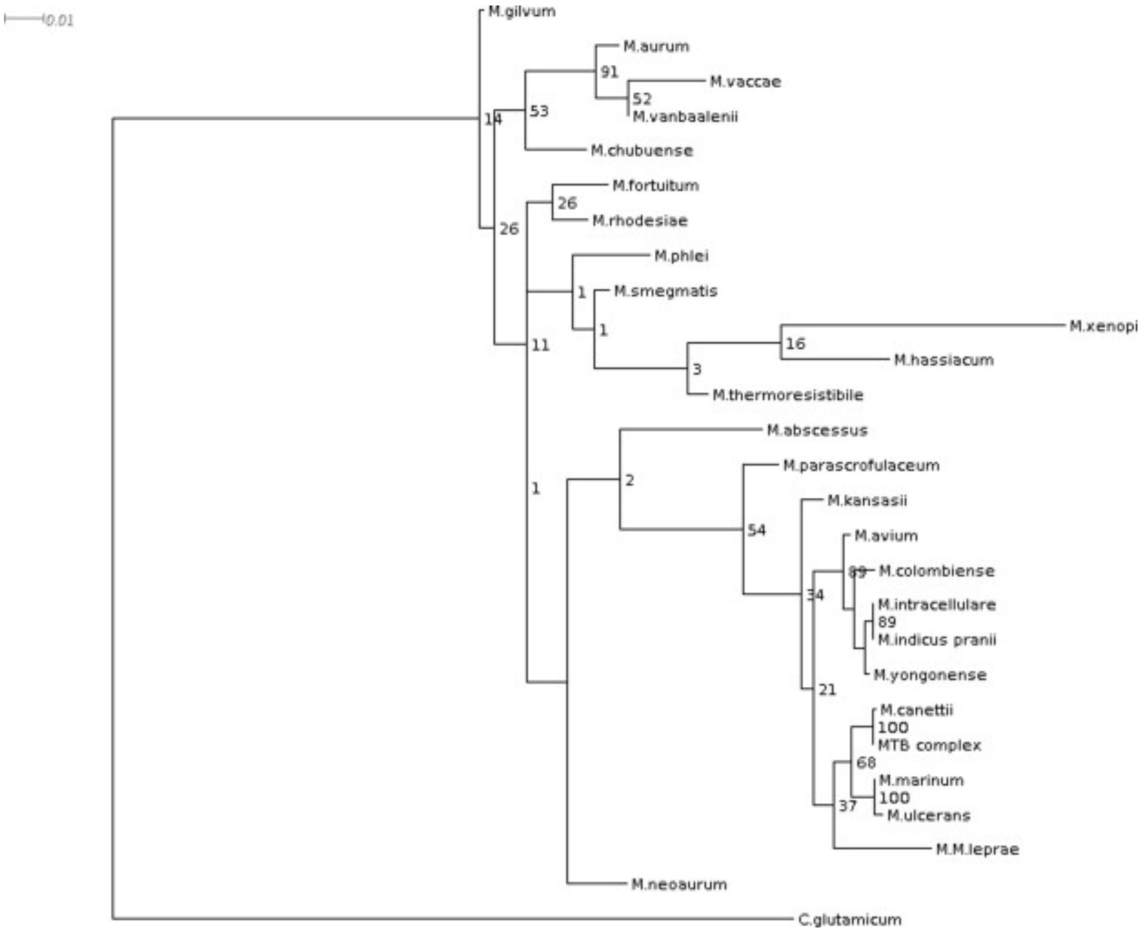
[39] C. Vilcheze, W.R. Jacobs Jr., Resistance to isoniazid and ethionamide in myocbacterium tuberculosis: genes, mutations and causalities, in: G.F. Hatfull, W.R. Jacobs Jr. (Eds.), Molecular Genetics of Mycobacteria, 2nd ed., ASM Press, USA, 2014.

[40] F. Bardou, A. Quemard, M.A. Dupont, C. Horn, G. Marchal, M. Daffe, Effects of isoniazid on ultrastructure of *Mycobacterium aurum* and *Mycobacterium tuberculosis* and on production of secreted proteins, Antimicrob. Agents Chemother. 40 (1996) 2459–2467.

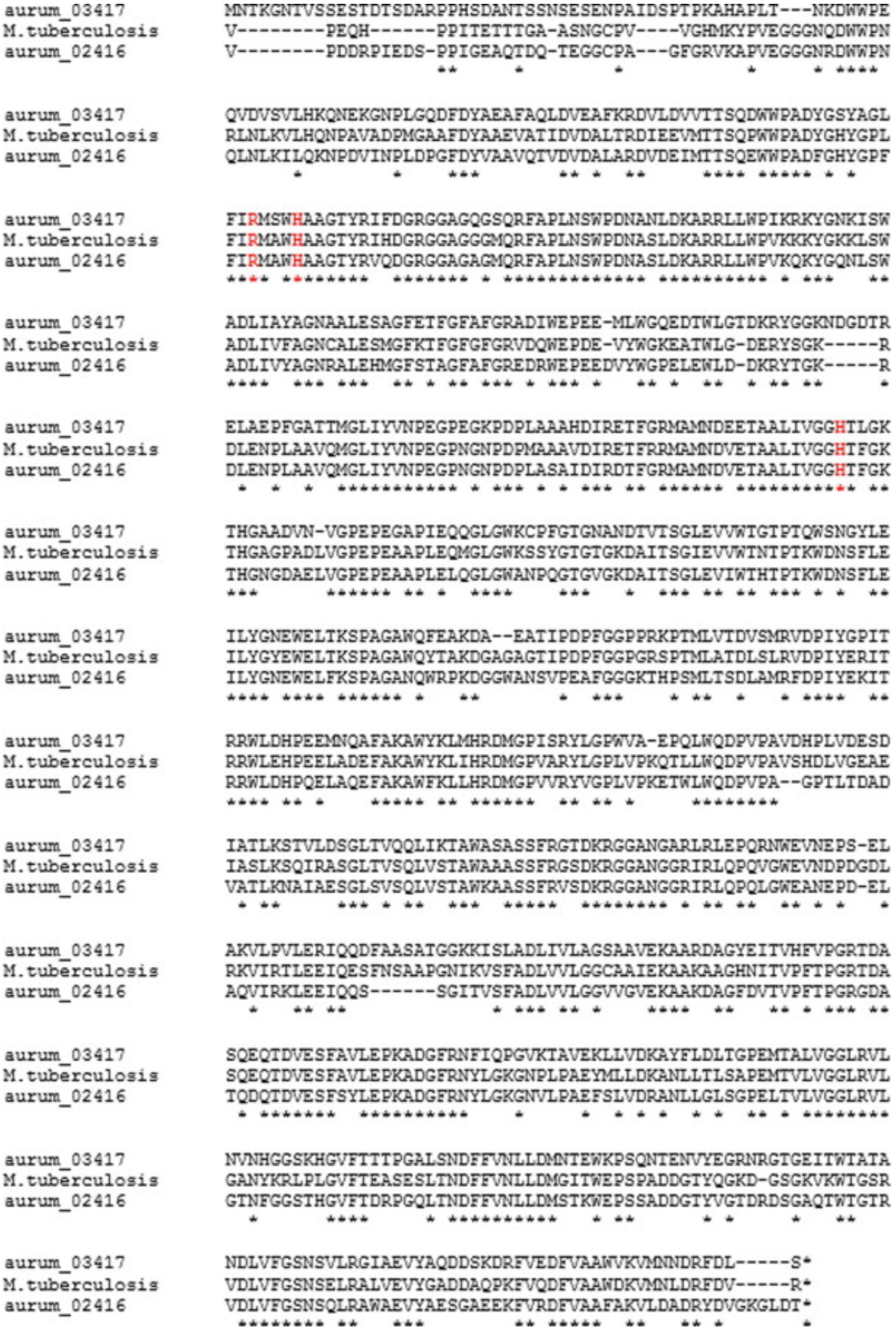
[41] M.C. Menendez, J.A. Ainsa, C. Martin, M.J. Garcia, KatGI and katGII encode two different catalases-peroxidases in *Mycobacterium fortuitum*, J. Bacteriol. 179 (1997) 6880–6886.

Supplementary Figure 1

***M. aurum* and the mycobacterium phylogeny[*] constructed using 16S rRNA sequences, which is less precise and with lower bootstrap support than that attained with whole genome sequence data (c.f. Fig. 2 in the main manuscript). [*]Constructed using *RaXML* and statistic support for lineages was based on 1000 bootstrap samples.**



Supplementary Figure 2
Comparative sequence of putative *katG* genes in *M. aurum* and *M. tuberculosis*.



```
aurum_03417         MNTKGNTVSSESTDTSDARPPHSDANTSSNSESENPAIDSPTPKAHAPLT---NKDWWPE
M.tuberculosis      V--------PEQH------PPITETTTGA-ASNGCPV----VGHMKYPVEGGGNQDWWPN
aurum_02416         V--------PDDRPIEDS-PPIGEAQTDQ-TEGGCPA---GFGRVKAPVEGGGNRDWWPN
                                       **     *        *           *     * ****

aurum_03417         QVDVSVLHKQNEKGNPLGQDFDYAEAFAQLDVEAFKRDVLDVVTTSQDWWPADYGSYAGL
M.tuberculosis      RLNLKVLHQNPAVADPMGAAFDYAAEVATIDVDALTRDIEEVMTTSQPWWPADYGHYGPL
aurum_02416         QLNLKILQKNPDVINPLDPGFDYVAAVQTVDVDALARDVDEIMTTSQEWWPADFGHYGPF
                          *        *    ***       ** *  **     **** ***** * *

aurum_03417         FIRMSWHAAGTYRIFDGRGGAGQGSQRFAPLNSWPDNANLDKARRLLWPIKRKYGNKISW
M.tuberculosis      FIRMAWHAAGTYRIHDGRGGAGGGMQRFAPLNSWPDNASLDKARRLLWPVKKKYGKKLSW
aurum_02416         FIRMAWHAAGTYRVQDGRGGAGAGMQRFAPLNSWPDNASLDKARRLLWPVKQKYGQNLSW
                    **** ********  ******* * ************* ********** * ***   **

aurum_03417         ADLIAYAGNAALESAGFETFGFAFGRADIWEPEE-MLWGQEDTWLGTDKRYGGKNDGDTR
M.tuberculosis      ADLIVFAGNCALESMGFKTFGFGFGRVDQWEPDE-VYWGKEATWLG-DERYSGK-----R
aurum_02416         ADLIVYAGNRALEHMGFSTAGFAFGREDRWEPEEDVYWGPELEWLD-DKRYTGK-----R
                    ****  *** ***  ** * ** *** * *** *   ** *  **  * ** **     *

aurum_03417         ELAEPFGATTMGLIYVNPEGPEGKPDPLAAAHDIRETFGRMAMNDEETAALIVGGHTLGK
M.tuberculosis      DLENPLAAVQMGLIYVNPEGPNGNPDPMAAAVDIRETFRRMAMNDVETAALIVGGHTFGK
aurum_02416         DLENPLAAVQMGLIYVNPEGPNGNPDPLASAIDIRDTFGRMAMNDVETAALIVGGHTFGK
                     *  *  *  *********** * *** * * *** ** ****** *********** **

aurum_03417         THGAADVN-VGPEPEGAPIEQQGLGWKCPFGTGNANDTVTSGLEVVWTGTPTQWSNGYLE
M.tuberculosis      THGAGPADLVGPEPEAAPLEQMGLGWKSSYGTGTGKDAITSGIEVVWTNTPTKWDNSFLE
aurum_02416         THGNGDAELVGPEPEAAPLELQGLGWANPQGTGVGKDAITSGLEVIWTHTPTKWDNSFLE
                    ***      ****** ** *  ****    ***   *  *** ** ** *** * *  **

aurum_03417         ILYGNEWELTKSPAGAWQFEAKDA--EATIPDPFGGPPRKPTMLVTDVSMRVDPIYGPIT
M.tuberculosis      ILYGYEWELTKSPAGAWQYTAKDGAGAGTIPDPFGGPGRSPTMLATDLSLRVDPIYERIT
aurum_02416         ILYGNEWELFKSPAGANQWRPKDGGWANSVPEAFGGGKTHPSMLTSDLAMRFDPIYEKIT
                    **** **** ****** *   **       *  ***    * **  *   * ****  **

aurum_03417         RRWLDHPEEMNQAFAKAWYKLMHRDMGPISRYLGPWVA-EPQLWQDPVPAVDHPLVDESD
M.tuberculosis      RRWLEHPEELADEFAKAWYKLIHRDMGPVARYLGPLVPKQTLLWQDPVPAVSHDLVGEAE
aurum_02416         RRWLDHPQELAQEFAKAWFKLLHRDMGPVVRYVGPLVPKETWLWQDPVPA--GPTLTDAD
                    **** ** *    ***** ** ******  ** ** *     ********

aurum_03417         IATLKSTVLDSGLTVQQLIKTAWASASSFRGTDKRGGANGARLRLEPQRNWEVNEPS-EL
M.tuberculosis      IASLKSQIRASGLTVSQLVSTAWAAAASSFRGSDKRGGANGGRIRLQPQVGWEVNDPDGDL
aurum_02416         VATLKNAIAESGLSVSQLVSTAWKAASSFRVSDKRGGANGGRIRLQPQLGWEANEPD-EL
                     * **     *** * **  ***  *****  ******** * ** **  ** * *   *

aurum_03417         AKVLPVLERIQQDFAASATGGKKISLADLIVLAGSAAVEKAARDAGYEITVHFVPGRTDA
M.tuberculosis      RKVIRTLEEIQESFNSAAPGNIKVSFADLVVLGGCAAIEKAAKAAGHNITVPFTPGRTDA
aurum_02416         AQVIRKLEEIQQS------SGITVSFADLVVLGGVVGVEKAAKDAGFDVTVPFTPGRGDA
                      *   ** **             * *** ** *    ****  **   ** * *** **

aurum_03417         SQEQTDVESFAVLEPKADGFRNFIQPGVKTAVEKLLVDKAYFLDLTGPEMTALVGGLRVL
M.tuberculosis      SQEQTDVESFAVLEPKADGFRNYLGKGNPLPAEYMLLDKANLLTLSAPEMTVLVGGLRVL
aurum_02416         TQDQTDVESFSYLEPKADGFRNYLGKGNVLPAEFSLVDRANLLGLSGPELTVLVGGLRVL
                     * *******  **********    *     *  * * *  * *  ** * ********

aurum_03417         NVNHGGSKHGVFTTTPGALSNDFFVNLLDMNTEWKPSQNTENVYEGRNRGTGEITWTATA
M.tuberculosis      GANYKRLPLGVFTEASESLTNDFFVNLLDMGITWEPSPADDGTYQGKD-GSGKVKWTGSR
aurum_02416         GTNFGGSTHGVFTDRPGQLTNDFFVNLLDMSTKWEPSSADDGTYVGTDRDSGAQTWTGTR
                      *      ****     * **********   * **      * *     *   **

aurum_03417         NDLVFGSNSVLRGIAEVYAQDDSKDRFVEDFVAAWVKVMNNDRFDL-----S*
M.tuberculosis      VDLVFGSNSELRALVEVYGADDAQPKFVQDFVAAWDKVMNLDRFDV-----R*
aurum_02416         VDLVFGSNSQLRAWAEVYAESGAEEKFVRDFVAAFAKVLDADRYDVGKGLDT*
                     ******** **   ***        ** *****  **   ** *       *
```

Supplementary Table 1
List of intergenic primers used for PCR amplification and Sanger sequencing serving to confirm the presence of duplicate/homologous *katG* and *embB* genes in *M. aurum*

Gene	Primer orientation	Primer sequence
katG1	Forward	GGAGATTTCCCGATCACAACCGTGATCACAG
	Reverse	CCGCTGATCAGTTCGAGACTGCACCCGTTC
katG2	Forward	CGACGAGGCCGAGGTCATCTACTGGGGC
	Reverse	CCCTACCGAATGTCGACGACAGCGCCGC
embB1	Forward	GCGCCCGACGCCGCCATCGAGGAAGG
	Reverse	CGGGGGTCTGGTCGAACAACGCGGTC
embB2	Forward	CCGACCATTGTGGAGCATCCCGACCCC
	Reverse	CGCCACCGACGTCTTCGAGATTCGTGAC

Supplementary Table 2
Number of orthologues between *M. aurum* and *M. tuberculosis* / *M. leprae*

Comparison with *M. aurum*	*M. tuberculosis*	*M. leprae*
No. of scaffolds with homology	10	13
No. of syntenous segments	67	73
No. of unique proteins	1,002	222
No. of unique *M. aurum* proteins	2,090	2,047
No. of orthologue clusters *	2,431	1,349
No. of 1-to-1 orthologues**	2,305 (94.8%)	1,299 (96.3%)

*Clusters of proteins with at least one representative of *M. aurum* and the other mycobacteria; ** orthologue clusters in which there is only one representative of each proteome.

Supplementary Table 3
Genes deemed essential for the survival of *M. tuberculosis* within a macrophage and their homologues in the surrogate species [1]

M. tuberculosis	*M. smegmatis*	*M. aurum*
mce1A (Rv0169)	*MSMEG_0134*	Present
mce1B (Rv0170)	*MSMEG_0135*	Present
mce1C (Rv0171)	*MSMEG_0136*	Present
mce1D (Rv0172)	*MSMEG_0137*	Present
lprK (Rv0173)	*MSMEG_0138*	Present
mceF (Rv0174)	*MSMEG_0139*	Present
Rv0175	*MSMEG_0140*	Present
Rv0176	*MSMEG_0141*	Present
Rv0177	*MSMEG_0142*	Present
Rv0178	*MSMEG_0143*	Present
sugC (Rv1238)	*MSMEG_5058*	Present
sugB (Rv1237)	*MSMEG_5059*	Present
sugA(Rv1236)	*MSMEG_5060*	Present
lpqY(Rv1235)	*MSMEG_5061*	-
pstA1(Rv0930)	*MSMEG_5780*	Present
pstC2(Rv0929)	*MSMEG_5781*	Present
pstS3(Rv0928)	*MSMEG_5782*	Present
espE (Rv3864)	-	-
eccA1 (Rv3868)	*MSMEG_0059*	Present
eccCa1 (Rv3870)	*MSMEG_0061*	Present
eccCb1 (Rv3871)	*MSMEG_0062*	Present
eccD1 (Rv3877)	*MSMEG_0068*	Present
hsaD (Rv3569c)	*MSMEG_6037*	Present
hsaA (Rv3570c)	*MSMEG_6038*	Present
Rv3552	*MSMEG_6003*	Present
Rv3551	*MSMEG_6002*	Present
echA20 (Rv3550)	*MSMEG_6001*	Present
fadE28 (Rv3544c)	*MSMEG_5994*	Present
Rv3542c	*MSMEG_5992*	Present
Rv3541c	*MSMEG_5991*	Present

[1] Rengarajan J, Bloom BR, Rubin EJ. Genome-wide requirements for *Mycobacterium tuberculosis* adaptation and survival in macrophages. *Proc Natl Acad Sci U S A* 2005,**102**:8327-8332.

Supplementary Table 4
Results from Sanger sequencing of putative *katG* and *embB* genes in *M. aurum*

Gene	Start codon identified	Stop codon identified	*Bam*HI site located
katG1 (2277bp)	Yes	Yes	Yes 737bp from start codon 729 bp from stop codon
katG2 (2235bp)	Yes	Yes	Yes 821bp from start codon
embB1 (3276bp)	Yes	Yes	No
embB2 (2985bp)	Yes	Yes	Yes 517bp from stop codon

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Jody Phelan
Principal Supervisor	Taane Clark
Thesis Title	A Bioinformatic analysis of *M. tuberculosis* and host genomic data

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?	BMC Genomics		
When was the work published?	Feburary 2016		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I selected all sequenced isolates from the global dataset described in chapter 4 based on a high depth of coverage. After this, I performed raw sequence data QC and removed all low-quality data. I evaluated several genome assembly programs on a subset of 20 strains to find the optimum program/settings. I chose Velvet based on its computational time requirement and assembly performance and performed genome assembly for all 518 strains. After this was complete, I applied several programs to improve the assembly. I aligned the assembled to the reference and performed phylogenetic reconstructions using subsets of the data (whole genome, PE/PPE). Using the alignment, I looked for evidence of positive selection and recombination. All analysis was performed using custom scripts I wrote in bash, R and python. I generated figures and tables for the manuscript in R. I co-wrote a first draft of the manuscript and incorporated comments from co-authors. Following this, I submitted to the

Student Signature: ______________________________ **Date:** ______________

Supervisor Signature: ______________________________ **Date:** ______________

Chapter 6

Recombination in *pe*/*ppe* genes contributes to genetic variation in *Mycobacterium tuberculosis* lineages

Phelan *et al. BMC Genomics* (2016) 17:151
DOI 10.1186/s12864-016-2467-y

BMC Genomics

RESEARCH ARTICLE Open Access

Recombination in *pe/ppe* genes contributes to genetic variation in *Mycobacterium tuberculosis* lineages

Jody E. Phelan[1], Francesc Coll[1], Indra Bergval[2], Richard M. Anthony[2], Rob Warren[3], Samantha L. Sampson[3], Nicolaas C. Gey van Pittius[3], Judith R. Glynn[4], Amelia C. Crampin[4,5], Adriana Alves[6], Theolis Barbosa Bessa[7], Susana Campino[1], Keertan Dheda[8,9], Louis Grandjean[1,10], Rumina Hasan[11], Zahra Hasan[11], Anabela Miranda[6], David Moore[1], Stefan Panaiotov[12], Joao Perdigao[13], Isabel Portugal[13], Patricia Sheen[10], Erivelton de Oliveira Sousa[7], Elizabeth M. Streicher[3], Paul D. van Helden[3], Miguel Viveiros[14], Martin L. Hibberd[1], Arnab Pain[15], Ruth McNerney[1] and Taane G. Clark[1,4*]

Abstract

Background: Approximately 10 % of the *Mycobacterium tuberculosis* genome is made up of two families of genes that are poorly characterized due to their high GC content and highly repetitive nature. The PE and PPE families are typified by their highly conserved N-terminal domains that incorporate proline-glutamate (PE) and proline-proline-glutamate (PPE) signature motifs. They are hypothesised to be important virulence factors involved with host-pathogen interactions, but their high genetic variability and complexity of analysis means they are typically disregarded in genome studies.

Results: To elucidate the structure of these genes, 518 genomes from a diverse international collection of clinical isolates were *de novo* assembled. A further 21 reference *M. tuberculosis* complex genomes and long read sequence data were used to validate the approach. SNP analysis revealed that variation in the majority of the 168 *pe/ppe* genes studied was consistent with lineage. Several recombination hotspots were identified, notably *pe_pgrs3* and *pe_pgrs17*. Evidence of positive selection was revealed in 65 *pe/ppe* genes, including epitopes potentially binding to major histocompatibility complex molecules.

Conclusions: This, the first comprehensive study of the *pe* and *ppe* genes, provides important insight into *M. tuberculosis* diversity and has significant implications for vaccine development.

Background

Tuberculosis disease (TB) is a major global public health problem, with control becoming difficult due to increasing drug resistance and in some populations HIV co-infection [1]. The available vaccine, Bacillus Calmette–Guérin (BCG), has limited efficacy and recent attempts to develop more effective protective vaccines have not been successful [2]. TB is caused by bacteria of the *Mycobacterium tuberculosis* complex, which have low overall genetic diversity and a striking clonal population structure. *M. tuberculosis sensu stricto* consists of seven lineages, including four that are predominant; 1 Indo-Oceanic, 2 East-Asian including Beijing, 3 East-African-Indian, 4 Euro-American [3]. These lineages are postulated to have differential impacts on pathogenesis, disease outcome and vaccine efficacy [4–7]. For example, modern lineages, such as Beijing and Euro-American Haarlem strains exhibit more virulent phenotypes compared to ancient lineages, such as East African Indian[8]. Whilst some genetic differences between lineages have been identified[3], the molecular mechanisms responsible for differences in pathogenesis and virulence remain largely unknown [8].

* Correspondence: taane.clark@lshtm.ac.uk
Ruth McNerney and Taane G Clark are joint authors
[1]Department of Pathogen Molecular Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, WC1E 7HT London, UK
[4]Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, WC1E 7HT London, UK
Full list of author information is available at the end of the article

Two groups of proteins, the PE and PPE families have been implicated in immune evasion and virulence [9]. Members of the *pe/ppe* gene families are characterized by the presence of proline-glutamate (PE) and proline-proline-glutamate (PPE) signature motifs near the N-terminus of their gene products [10]. The *pe* (99 loci) and *ppe* (69) gene families constitute ~7–10 % of the coding potential of *M. tuberculosis* and are scattered throughout the genome [9]. The families can be subdivided based on similarities in their N-terminal regions [11]. Many of the *pe* and *ppe* gene products are predicted to be localised to the cell membrane or secreted including those in the PE_PGRS domain containing subgroup and the PPE_MPTR domain containing subgroup [12, 13]. It has been speculated that these proteins may play a role in virulence [14]. *Pe/ppe* genes are differentially expressed during infection [15] and some PE/PPE proteins have been shown to elicit immune responses by the host [14, 16] and there is evidence that the PGRS domain can inhibit antigen processing [16, 17].

Whilst *pe_pgrs* and *ppe_mptr* genes represent some of the most variable *M. tuberculosis* regions, some members of the *pe/ppe* family are conserved across strains and species, therefore implying different functional roles. Only the protein structures of PE25 and PPE41 have been characterised [18], and in lieu of experimental and functional work, insights into their function and interaction partners must come from in silico analysis of large-scale ‘omics data. However, due to the repetitive nature and high GC content genetic variation in the *pe/ppe* genes, it has been difficult to characterize them using traditional mapping approaches, leading to their systematic exclusion from analysis [18]. There have been conflicting studies reporting either high or little or no sequence divergence [19–21], but studies have been limited by the number of genes and diversity of strains analysed.

There is a need to fully characterize *pe/ppe* family sequence diversity across strain-types to provide better understanding of these genes and their possible role in virulence and immune evasion. The availability of high throughput short sequencing technologies has revolutionized the study of *M. tuberculosis* genetic diversity. In an attempt to characterize these elusive genes we have performed whole genome assembly on next generation sequence data with a high depth of coverage across the *pe/ppe* gene regions from 518 clinical and experimental isolates. These isolates represent the four major lineages, each with known informative barcoding SNPs [3]. The approach was validated by examination of 21 reference genomes from established databases (www.tbdb.org; www.ebi.ac.uk), including 2 new strains with complete genomes sequenced using long read Pacific Bioscience (PacBio) technology [22, 23].

Results

Assembly of *M. tuberculosis* genomes

Conventional alignment-based analysis approaches have been of limited use in analysis of highly repetitive loci, including the *pe/ppe* genes. Here, we *de novo* assembled the genomes of 518 samples from 9 different countries covering the four main lineages (1 (n = 42), 2 (n = 38), 3 (n = 53), and 4 (n = 385)), with high sequence coverage in *pe/ppe* genes (mean 233-fold, range 100–1544) (Additional file 1: Tables S1 and S2). For each sample, at least 120 of the 168 *pe/ppe* genes were fully assembled and at least 90 % assembled for the remaining 48 genes (Additional file 1: Table S3). This level of assembly quality ensured low levels of assembly fragmentation and minimised poor gene characterization. Subsequent analysis involving manual inspection or re-mapping of reads to the assemblies using REAPR software, revealed all genes (168 *pe/ppe*; 3,654 other genes; 2,820 with an assigned function) to be of high quality (median REAPR score of 1 across all bases, reflecting high levels of accuracy in genome assemblies). A further 21 independent complete reference genomes representing all four lineages (Additional file 1: Table S1), were aligned against H37Rv to call variants, and used to further validate the results found in the assembled dataset.

Variant detection and population genetic analysis

A total of 50,539 genome-wide SNPs were identified by comparing the 518 assembled genomes to the H37Rv (lineage 4, Euro-American T) reference strain. Of these, 5,853 (11.6 %) SNPs were located within *pe/ppe* regions, with greater density than the rest of the genome (median SNPs per kb: *ppe/pe* = 12.9, non-*ppe/pe* = 9.1, Wilcoxon $P < 2.2 \times 10^{-14}$). In the 257 Malawi samples, our assembly procedure revealed 3,467 additional SNP variants genome-wide (1,438 (41.5 %) SNPs in 72 *pe/ppe* genes) compared to the standard approach of aligning short reads to the H3Rv reference. Of the 50,539 SNPs inferred from the assemblies, the majority (45,681, 90.3 %) were located in coding regions from all genes and consisted of 28,235 (61.8 %) non-synonymous SNPs and 17,446 (38.2 %) synonymous SNPs. This observation is in agreement with the higher abundance of non-synonymous mutations reported in the literature [19]. A large number of rare variants (i.e. present in only one isolate) were observed in all lineages, indicative of purifying selection and population expansion described by others [24]. The peaks in the spectrum represent a number of SNPs that are fixed in all isolates from sub-lineages (Additional file 2: Figure S1).

The ratio of non-synonymous to synonymous mutations was similar in *pe/ppe* and other genes (median: *pe/ppe* genes = 1.65, other genes = 1.75, Wilcoxon $P =$ 0.68). The density of non-synonymous mutations was

2.98 times greater in *pe/ppe* genes compared to others (*pe/ppe* genes: 1 every 3933 bp, other genes: 1 every 11,706 bp, Wilcoxon $P < 0.0001$), consistent with another report [25]. When analysed by sub-family we observed the greatest ratio of densities in the *pe_pgrs* genes (*pe_pgrs* 3.89) compared to the other types (*ppe* 1.75, *pe* (non-*pe_pgrs*) 1.80), similar to that reported previously [25]. The nucleotide diversity (π) was ~2-fold greater in the *pe/ppe* genes (median: *pe/ppe* genes 2.7×10^{-4}, other genes 1.4×10^{-4}, Wilcoxon $P < 1.4 \times 10^{-10}$). Although estimates of genetic diversity may be influenced by sampling bias, nucleotide diversity varied by lineage, being greater in lineage 1 (Indo-Oceanic median: *pe/ppe* 1.7×10^{-4}, other 9.0×10^{-5}) and lower in lineage 2 (East-Asian median: *pe/ppe* 7.3×10^{-5}, other 0) (Additional file 1: Table S2), all consistent with previous work [3]. Loci identified as being highly diverse ($\pi > 0.003$, top 0.2 %, Table 1, Fig. 1), included 5 *pe/ppe* genes (*pe_pgrs3, pe_pgrs4, ppe57, ppe59* and *ppe60*), and 3 others *(Rv0030, Rv0095c and lppB*). The diversity per gene was compared to those from 21 complete reference genomes, and peaks were observed at *Rv0095c, pe_pgrs3, pe_pgrs4, ppe57 and ppe60,* independently supporting five out of the eight loci identified in the 518 global samples (Additional file 3: Figure S2).

Phylogenetics

To examine the link between genetic variation and lineage, a phylogenetic tree was constructed using the 50,539 SNPs. It revealed clustering by lineage, thereby further validating the quality of the assembled genomes (Additional file 4: Figure S3). However, a similar analysis using 5,853 *pe/ppe* specific SNP positions led to a tree with lineage 2 being split into two distinct clades, surrounded by lineage 4 strains (Fig. 2a). Subsequent analysis using SNP-based population differentiation F_{ST} and site-specific log likelihood scores approaches (Additional file 5: Figure S4) revealed that the *pe_pgrs3* gene (genomic position 333 kb, lineage 2 – 104 SNPs differentiating) was predominantly responsible for the ambiguity. Removal of the 281 SNPs in the *pe_pgrs3* gene led to a *pe/ppe*-based tree that clustered by lineage (Fig. 2b), very similar in topology to that based on the genome-wide SNPs (Additional file 4: Figure S3). This demonstrated that a core set of *pe/ppe* SNPs appears to be lineage specific, and further analysis revealed a set of 87 (1.4 %) SNPs (66 non-synonymous) that were lineage specific, potentially forming the basis of a lineage-specific molecular barcode (Additional file 1: Table S4). None of these 87 mutations were present in *M. bovis* (GCA_000195835) or *M. africanum* (NC_015758.1) sequences, and therefore

Table 1 Loci that are highly diverse, with recombination, or under selective pressure

Gene	Locus	No. SNPs	Diversity *π*	*phi p-value*	*phi p*-value[a]	*dN/dS (w)*	No. sites[b]	Lineage specific *phi*
Rv0030	*Rv0030*	3	**0.0033**	1.000	1.000	-	0	-
Rv0095c	*Rv0095c*	10	**0.0059**	**0.005**	**0.021**	10.13	3	-
Rv0182c	*sigG*	3	0.0003	**0.046**	**0.046**	-	0	-
Rv0278c	*pe_pgrs3*	130	**0.0193**	**<0.001**	**<0.001**	10.5	49	1,3,4
Rv0279c	*pe_pgrs4*	49	**0.0035**	**0.001**	0.419	10.5	20	-
Rv0282	*eccA3*	5	0.0005	**0.007**	0.210	9.697	6	-
Rv0850	*Rv0850*	2	**0.0031**	1.000	1.000	9.264	4	-
Rv0978c	*pe_pgrs17*	9	0.0005	**0.003**	1.000	10.495	9	-
Rv1148c	*Rv1148c*	18	0.0022	**<0.001**	0.015	10.492	5	4
Rv1793	*esxN*	6	0.0023	**0.034**	0.159	9.694	2	4
Rv1945	*Rv1945*	18	0.0010	**<0.001**	**0.026**	10.433	5	-
Rv2048c	*pks12*	80	0.0008	**<0.001**	**0.012**	10.5	79	4
Rv2543	*lppA*	8	0.0015	**0.006**	**0.002**	10.036	5	4
Rv2544	*lppB*	60	**0.0123**	**<0.001**	**<0.001**	5.336	33	1,2,4
Rv3425	*ppe57*	31	**0.0154**	0.431	1.000	10.5	21	-
Rv3429	*ppe59*	19	**0.0041**	**<0.001**	0.084	10.419	29	4
Rv3466	*Rv3466*	6	0.0010	**0.004**	0.373	7.757	3	-
Rv3478	*ppe60*	105	**0.0061**	**<0.001**	**0.004**	7.502	54	4
Rv3619c	*esxV*	3	0.0022	**0.025**	1.000	10.391	2	-

π nucleotide diversity, *phi* recombination, *NS* not significant

[a] after removing sites under selection

[b] number of sites under selection using the Bayes Empirical Bayes method

Bolded refers to $\pi > 0.003$ *or phi p*-value < 0.05

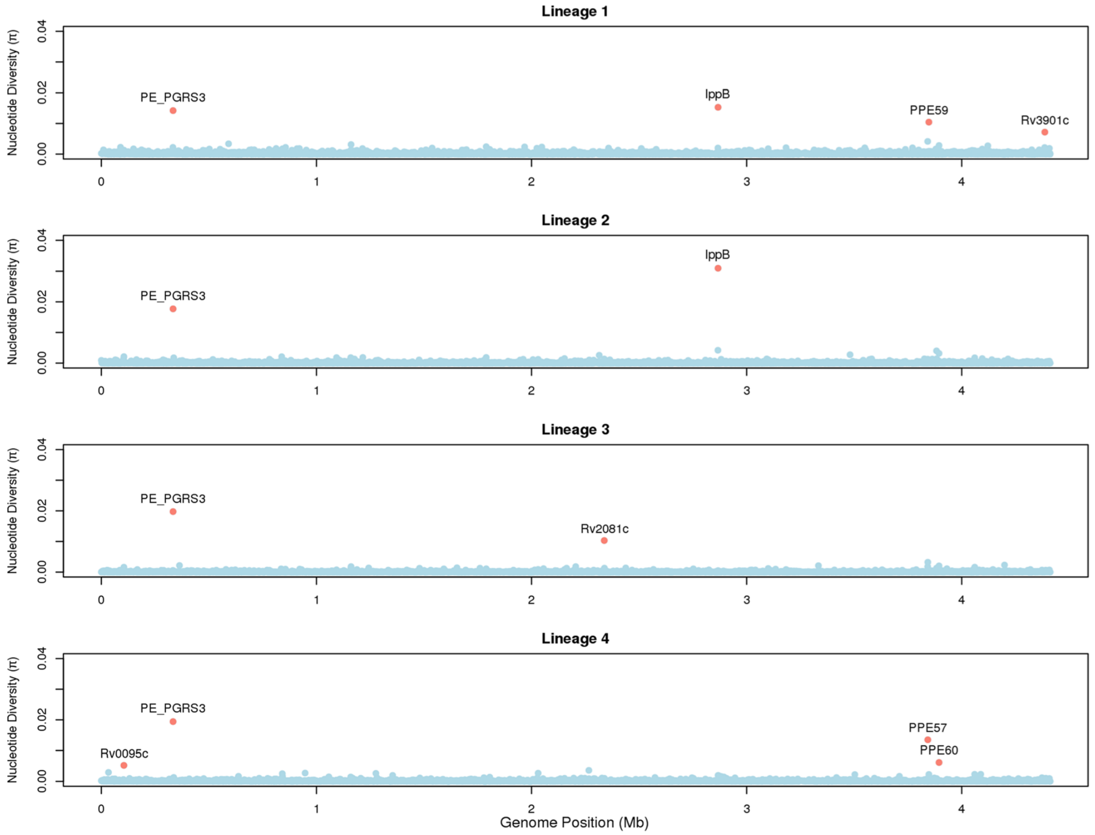


Fig. 1 Nucleotide diversity (*π*) across the genome by lineage; genes with high diversity (*π* > 0.003) are highlighted. The *pe_pgrs3* gene appears to have high nucleotide diversity in all lineages. Some lineage-specific hotspots are seen in lineage 1 (*ppe59* and *Rv3901c*), lineage 3 (*Rv2081c*) and lineage 4 (*ppe57* and *ppe60*)

robust as *M. tuberculosis* lineage-specific markers. Using only the *pe_pgrs3* SNPs led to a tree with two large clades (Additional file 6: Figure S5), one containing H37Rv and strains with similar sequence, and the other consistent with isolates similar to M323 and 18b strains (Additional file 1: Table S1b), both undergoing recent sequencing using PacBio long read technology. The M323, 18b and similarly clustered assembled samples have a *pe_pgrs3* gene with conserved regions at both 3′ and 5′ ends, surrounding a highly similar hypervariable core. A different hypervariable core is present in H37Rv and similarly clustered assemblies, which interestingly is also present in the *pe_pgrs4* gene of 18b, and recombination is a potential explanation.

Recombination detection

Although it has been thought that *M. tuberculosis* undergoes little or no homologous recombination, PE_PGRS and PPE_MPTR families contain long domains comprised of series of tandem repeats, giving them a higher propensity to undergo recombination. There is published evidence of intra-chromosomal cross-over ahead of a few loci [9], including *pe_pgrs3, pe_pgrs4,* and *ppe1* [26]. We hypothesized that recombination may be the reason for the observed high genetic diversity and distortion in the *pe/ppe* tree. We applied the pairwise homoplasy index (*phi*) method [27] genome-wide to establish if there was any evidence of recombination in *pe_pgrs3* and other loci (Fig. 3). The method calculates a *p*-value (*phi* P) of observing the sequence data under the null hypothesis of no recombination. The analysis revealed 16 genes with potential recombination events (*phi* $P < 0.05$) present across all lineages: 5 in *pe/ppe* genes (*pe_pgrs3, pe_pgrs4, pe_pgrs17, ppe59 and ppe60*), and 11 others (*Rv0095, sigG, eccA3, Rv1148, esxN, Rv1945, pks12, lppA, lppB, Rv3466* and *esxV*).

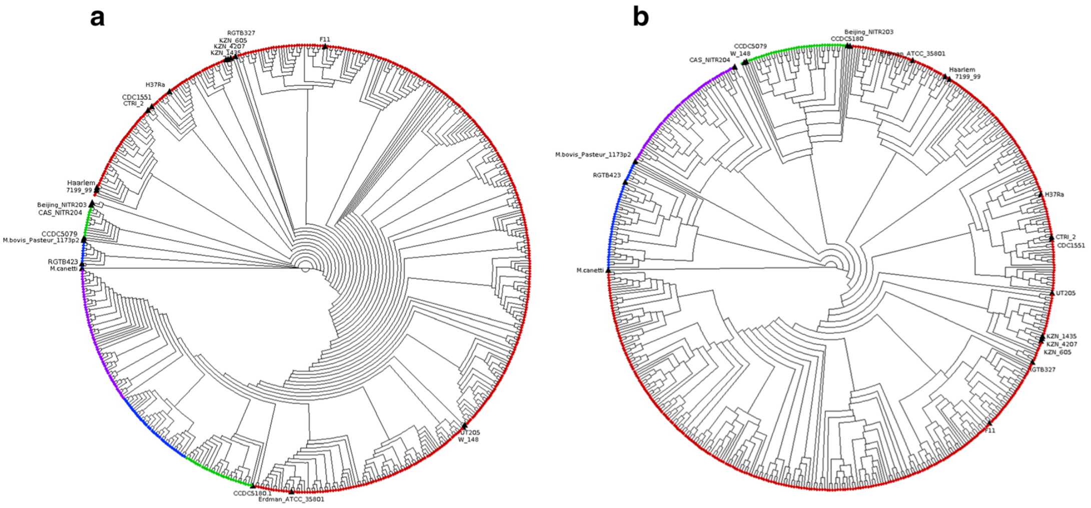


Fig. 2 Phylogenetic tree constructed with SNPs. **a** *pe/ppe* genes (5,404 SNPs, 10 % of the genome); the *pe_pgrs3* gene was identified as having SNPs leading to the lineages not perfectly clustering, potential evidence of recombination affecting these loci. **b** *pe/ppe* genes excluding *pe_pgrs3* (5,572 SNPs, ~10 % of the genome). Clear clustering according to lineage can be seen (Lineage 1 (Indo-Oceanic, *green*), lineage 2 (East-Asian (Beijing), *blue*), lineage 3 (East-African-Indian, *purple*), lineage 4 (Euro-American, *red*)). Reference genomes are labelled. *M. canetti* is annotated in cyan

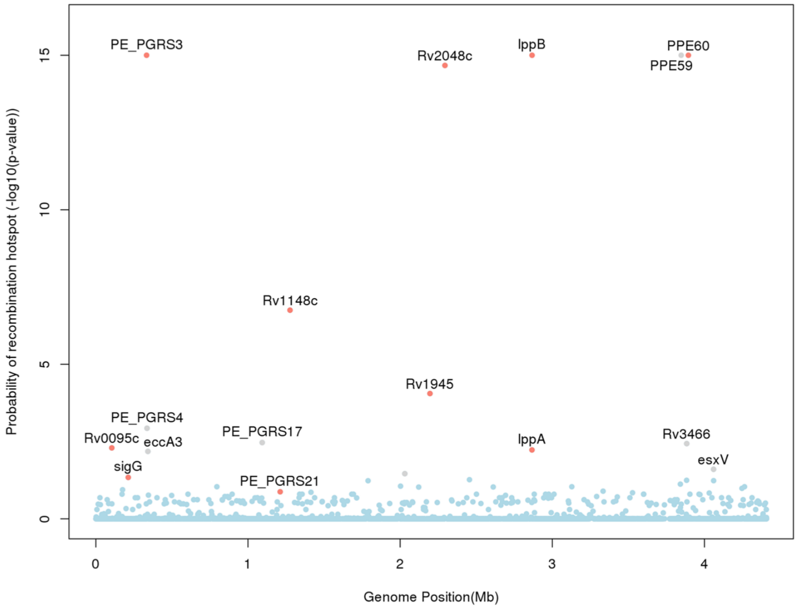


Fig. 3 Evidence of recombination at a gene level. A Manhattan plot showing genes that are likely to be recombination hotspots. The (−log10) *p*-value for recombination is plotted against genome position. All genes with *p*-values <0.05 are labelled. Genes labelled in colour grey *(eccA3, pe_pgrs4, pe_pgrs17, ppe59, Rv3466 and esxV)* become statistically non-significant after removing sites under selection

It could be expected that the vast majority of any genomic recombination events are intra-lineage and that these events will pass unnoticed by other analyses, especially in studies of small sample size. Lineage-specific hotspots were also present (Additional file 7: Figure S6), including possible pathogenicity factors *lppA/lppB* in lineage 2 (Beijing) and *pe_pgrs3* in lineage 4. An analysis of the 21 complete reference genomes revealed an overall high degree of concordance of the homoplasy *phi* statistic with the assembled data, with six recombination peaks in common (*Rv0095c, pe_pgrs3, pe_pgrs4, pe_pgrs17, Rv1148c* and *Rv1945*) (Additional file 8: Figure S7). Together, these results provide evidence for recombination.

Detecting selection pressure

It is possible that recombination and population expansion [24] could have introduced not only the observed increased diversity in the *pe/ppe* genes, but contributed to an excess of non-synonymous mutation diversity in general; especially in genes expected to be under positive or diversifying selection such as the cell wall component genes [24]. Proteins in contact with the host proteome could be under pressure to change their amino acid sequence in order to avoid detection or unfavourable interaction with the host immune system. We decided to investigate the role of selection in the *pe/ppe* genes compared to other categories of genes. The distribution of *dN/dS* values (denoted ω, = 1 neutral evolution, >1 positive selection, <1 purifying selection), calculated for each gene across all sites and branches of the phylogenetic tree, was similar between *pe/ppe* and other genes (median ω: *pe/ppe* genes 0.81, other genes 0.73; Wilcoxon $P = 0.16$). These values are broadly similar to those previously reported on much lower numbers of samples and *pe/ppe* genes [25]. The genes were further divided into functional Clusters of Orthologous Groups (COG) categories [28]. Higher median ω values were observed in genes associated with signal transduction mechanisms (median = 0.95), perhaps due to their contact with the host, and the lowest values found in genes associated with RNA processing and modification (median = 0.38) (Additional file 9: Figure S8).

In most genes it would be expected that only a small subset of sites would undergo positive selection and so calculation of a single ω value over all sites in the gene may dilute an effect. For example, this is possible in *pe/ppe* genes where there is less variation in the N- compared to the C-terminus [21]. We therefore used a likelihood ratio based approach that accounts for the variability of ω between sites. After implementation, we detected a greater proportion of *pe/ppe* loci under positive selection compared to other genes ($\omega > 1$ and $P < 0.05$: *pe/ppe* genes 65 (39 %) vs. other genes 590 (15 %)). This observation remained consistent when the non-*pe/ppe* genes were subdivided into functional categories (*P*-values for evidence of ω >1, Wilcoxon P <0.001) (Fig. 4). Using the COG categories, the genes associated with cell motility and the *pe/ppe* genes again showed greater evidence of significant positive selection (Additional file 10: Figure S9). All genes annotated as possible recombination hotspots were identified as being under positive selection, except *Rv0182c.* To localize the specific polymorphisms under selection we applied the Bayes Empirical Bayes (BEB) method [29], and identified a small number of sites in each gene (median (range): *pe/ppe* genes 0 (0–60), other genes 0 (0–48), $P = 1.2 \times 10^{-10}$). In total 99 *pe/*ppe genes had sites under positive selection, including ten genes with selection at more than ten sites (Additional file 1: Table S5). For 1,106 non-*pe/*ppe genes, only 37 had ten or more sites under positive selection. The proportion of segregating sites under positive selection (S_p/S_s) per gene was higher in the *pe/ppe* loci compared to others (*pe/ppe* genes 0.04, other genes 0.00, Wilcoxon $P = 2.58 \times 10^{-7}$). There was a correlation between the number of positively selected and segregating sites (Pearson's *r*, *pe/ppe* 0.81, and other genes 0.32).

We considered the 3,686 sites in the 1,106 non-*pe/ppe* genes with some evidence of positive selection ($\omega > 1$). These sites were compared to a list of drug resistance-conferring mutations (www.tbdb.org), which because of a survival advantage may be expected to be under positive selection. Eighteen drug resistance markers were found, including in *inhA* (I21T, S94A, I194T, P251A; associated with the drug isoniazid), *katG* (S315T; isoniazid), *gyrA* (A90V; fluoroquinolones), *rpoB* (P45L, rifampicin), *rpoL* (K43R; rifampicin), and *ponA1* (P631S; rifampicin). Other regions of interest included *rodA* (T336S) involved in cell wall processes and required for survival in primary murine macrophages, and *pks6* (V504L) involved in lipid metabolism and in vitro growth. Repeating the recombination detection analysis after removing the sites under positive selection identified by the BEB method, revealed six genes that lost their statistical significance (*phi* $P > 0.05$, *eccA3, pe_pgrs4, pe_pgrs17, ppe59, Rv3466 and esxV*), leaving 10 as crossover hotspots (Fig. 3). Given that variation in these genes is not caused by positive selection it is highly likely that recombination hotspots are indeed present at these ten loci. The proportion of sites under selection was high for *lppA* (7 %) and *lppB* (43 %) loci. The BEB method identified 38 codons in *lppA*/B at which $\omega > 1$, with almost all the related mutations present in lineage 2 (East-Asian) samples. None of these codons were in previously described conserved positions [30], implying that the core function of the protein was not disturbed, and the mutations may contribute to antigenic variation.

Selection on epitopes

Epitopes potentially binding to major histocompatibility complex molecules were predicted in all PE/PPE proteins using the netMHCpan software (Additional file 1:

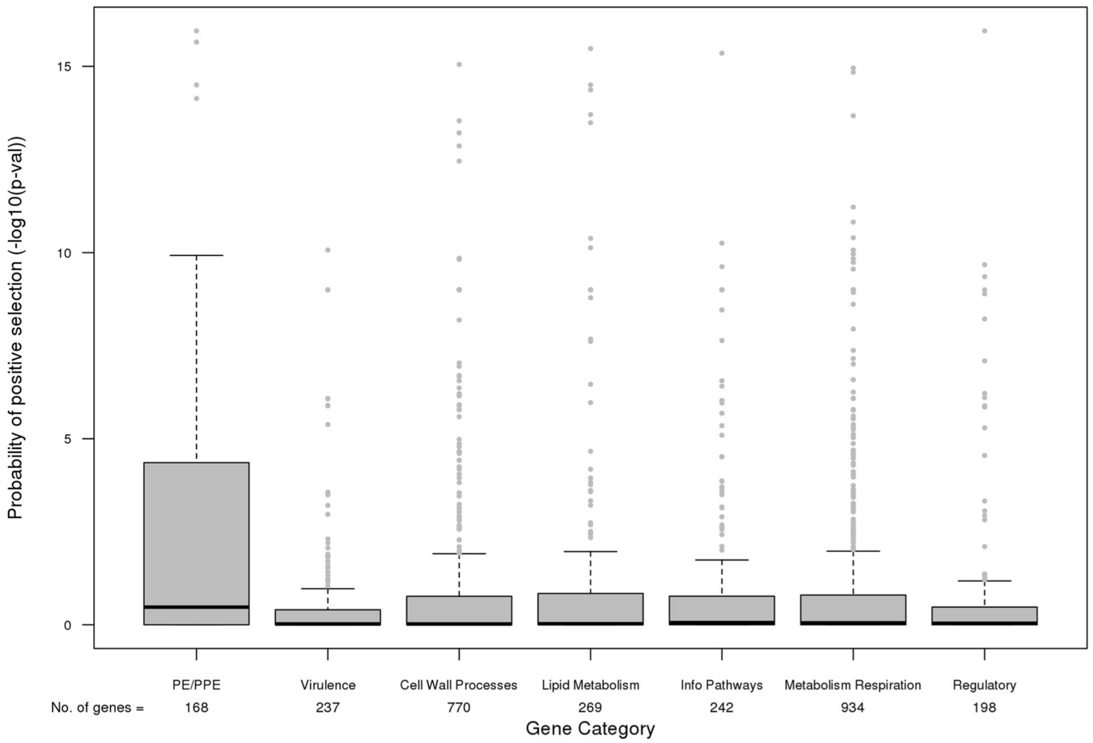


Fig. 4 Evidence of positive selection between the *pe/ppe* and other genes by functional annotation. Distributions of (–log10) *p*-values for positive selection (evidence of ω >1) across the *pe/ppe* and other genes by functional annotation

Table S6). The number of epitopes varied by *pe/ppe* gene (median 45, range 0 – 455). Some *pe/ppe* sites identified as being under selection using the BEB approach did overlap with regions predicted to be epitopes. In particular, for 10 genes (*pe6, pe_pgrs26, pe18, pe_pgrs49, pe_pgrs60, ppe27, ppe57, ppe59, ppe60 and ppe65*), more than 20 % of predicted epitopes had sites under positive selection (Additional file 1: Table S6).

Discussion

Members of the PE/PPE family of proteins have been found to trigger innate immune responses, are targets of the adaptive immune system, and potentially a rich source of diagnostic and vaccine antigens. As large 'omic studies in *M. tuberculosis* have often excluded *pe/ppe* genes from analysis (e.g. [3]), the understanding of their function and diversity is poor compared to other loci. Assessing diversity across *M. tuberculosis* strain types is critical, as lineages may vary in propensity to transmit and cause disease. By applying a *de novo* assembly approach, we were able to characterize accurately nearly all 168 *pe/ppe* genes in 518 isolates with high genomic coverage, representing lineages 1 (Indo-Oceanic), 2 (East-Asian), 3 (East-African-Indian) and 4 (Euro-American). After identifying ~50 k genome-wide SNPs from whole genome alignments, we confirmed that *pe/ppe* genes, especially the *pe_pgrs* family, have a high density of non-synonymous mutations compared to other *M. tuberculosis* loci. This observation is consistent with their involvement in antigenic variation and immune evasion, where proteins that are directly exposed to host immune surveillance tend to show higher levels of polymorphism. A lower degree of polymorphism in the *ppe* genes (compared to *pe_pgrs*) is likely to reflect a strong functional constraint of the PPE proteins.

Using all SNPs in a phylogenetic analysis, we observed clustering by *M. tuberculosis* lineage and therefore consistency with other published topologies [3, 31]. There was evidence of lineage specific *pe/ppe* repertories, with a very similar phylogeny being attained by restricting analysis to all polymorphisms in 167 PE/PPE genes (excluding *pe_pgrs3*), as well as a derived subset of 87 informative SNPs. The *pe_pgrs3* gene had high nucleotide diversity across all lineages, was not lineage informative, and is likely to be have been subject to recombination in lineages 1, 3 and 4. Both *M. bovis* and *M. canetti* contain two genes annotated as orthologues of *pe_pgrs3*, providing further evidence towards the propensity of this region to undergo genomic rearrangements. Interestingly the positioning of the *M. tuberculosis* reference strains in the *pe/ppe* gene phylogenetic tree was altered; some strains clustering near the *M. canetti* and ancestral strains while some of the known virulent reference strains were positioned at a further distance. Further study is needed to elucidate this effect. Other recombination and diversity hotspots included *lppB* (lineages 1, 2, and 4) and *ppe60* (lineage 4) genes, both known to have undergone homologous recombination. *LppB* (and *lppA)* are non-essential exported lipoproteins that are unique to pathogenic mycobacteria and may encode antigens [32, 33]. The *lppA/B* SNPs driving this effect were found mostly in lineage 2 (Beijing) strains, and seemed to be conferring a selective advantage. The role of lppA/B proteins on virulence should be investigated further. Although, *pe_pgrs17*, whose protein is in contact with the host immune system [34], was identified as a recombination hotspot, this observation may be confounded by positive selection. However, recombination has been described in *pe_pgrs17*, with large numbers of SNPs and indels in the *pe_pgrs17* and *pe_pgrs18* pair observed across the different lineages, potentially arising from gene conversion events [35]. We can rule out the results being confounded due to a sampling frame that included different geographical regions, as there was strongest clustering by lineage and not geographical source.

Across all *M. tuberculosis* genomes there was evidence that most genes were undergoing purifying selection pressures ($dN/dS < 1$). However, the *pe/ppe* genes were most likely to be under positive selection ($dN/dS > 1$), consistent with some PE/PPE proteins providing antigenic variation. It is possible the *dN/dS* ratios may be underestimated, as the methodology is more appropriate to divergent species and not for comparisons within a population [25]. Further, the signatures from very localised regions of selection within a gene may be diluted by surrounding genetic variation. A site-specific analysis confirmed the results from the gene-based *dN/dS*. Whilst the majority of the sixty-five genes identified as being under positive selection had only a single positively selected site, a disproportionate number of *pe_pgrs* genes had multiple positively selected sites. A potential limitation of this analysis is the time dependence of *dN/dS* for closely related bacterial genomes. This leads to possible over-estimation of the *dN/dS* and difficulties in interpretation when comparing the strength of selection between genes, genomes or populations over very short time-scales [36]. The power of the *dN/dS* statistic to detect positive selection is reduced when samples come from a single population [25]. In addition to the site under selection, multiple neighbouring and linked sites may show evidence of selection due to hitchhiking effects.

Our findings provide potential insights into the use of PE/PPE proteins as vaccine components. The high levels of polymorphism observed and the lineage-specific nature in certain members of these protein families could limit their effectiveness. A PE/PPE protein that displays higher sequence conservation across many strains may be a more effective vaccine candidate. For example, the highly immunogenic PE_PGRS62 protein has been considered as a vaccine target [37], and as only one of the 14 non-synonymous mutations observed was lineage specific, it may have broad strain coverage. However, one roadblock is the limited immunogenicity data available at the *pe/ppe* epitope level. It has been found that human T-cell epitopes are highly conserved in the *M. tuberculosis* complex [38], and like others [25] we found many epitopes predicted in PE/PPE proteins. Our analysis revealed a number of *pe/ppe* genes with a high proportion of epitopes potentially subjected to diversifying or positive selection. As these epitopes may be used by *M. tuberculosis* to evade the host immune system they would be relevant for TB vaccination strategies.

A cohesive understanding of the function of the 168 PE/PPE family of proteins remains elusive. By analysing SNP variation in 518 samples across the four main *M. tuberculosis* lineages we identified *pe/ppe* genes that are highly diverse, recombination hotspots and under positive selection. Such analyses can assist with prioritising candidates for functional studies, potentially leading to TB control measures, such as vaccines, diagnostics and drugs.

Conclusions

Human tuberculosis poses a major burden on health services worldwide. There is a need to understand the complex interactions between the human host and bacterial pathogen so that new control measures, such as vaccines and drugs, can be developed. Recent technological advances have allowed large-scale studies to determine the genetic signatures of strain-types or ancestral lineages and drug resistance outcomes. Despite this advance, some highly variable regions of the genome are often excluded [39, 40]. This includes the *pe/ppe* gene family, whose members are thought to interact with the human immune system, but little is still known of their diversity and function. Here we present the first comprehensive study of the genetic diversity of the 168 *pe/ppe* genes. We find most genes vary in a lineage specific manner, consistent with strain-specific repertories. However, there were exceptions to this pattern, with evidence of some genes undergoing genetic cross-over events. Further, by looking for the genes under selective pressure genome-wide, we found enrichment in the number of *pe/ppe* genes undergoing positive selection. Overall, our work highlights the importance of *pe/ppe* genes, describes their suitability as vaccine candidates, and provides the basis for further exploration of the proteins involved in the host immune system and pathogen interactions.

Methods

The raw sequencing fastq files for 518 *M. tuberculosis* samples with more than 100-fold genomic coverage were sourced from the PolyTB [41], rapid TB [42] and global drug resistance (Coll F, McNerney R, Hill-Cawthorn G et al. Whole genome association analysis of a global collection of Mycobacterium tuberculosis clinical isolates gives new insight into drug resistance, Submitted) projects (Additional file 1: Table S1a). A list of ENA accession numbers is available for download (http://pathogenseq.lshtm.ac.uk/ppe). Lineages were inferred using robust barcoding SNPs [3]. Lineages 1, 2, 3 and 4 were represented with 42, 38, 53 and 385 samples from each respectively (Additional file 1: Table S2). A separate set of twenty-one samples representing lineages 1 to 4 with complete or near complete genomes were used for validation (Additional file 1: Table S1b). In particular, all analyses performed on the main 518 samples were also applied to the validation dataset in an attempt to confirm signals and potentially rule out spurious findings. Assembly of all short reads was performed using MaSuRCA, SGA, Velvet and SPAdes [43–46] software, run in paired end mode with default and recommended parameters, across multiple k-mer values ranging from 31 to 91. The final Velvet run was implemented with a k-mer value of 63. Quast [47] software was used to extract assembly quality metrics using the H37Rv strain (Gene bank: AL123456) as the reference. The Samtools rmdup utility [48] was used to remove duplicates from each sample's BAM file, and picard SamToFastq (http://broadinstitute.github.io/picard/) was used to convert the BAM files to fastq format. IMAGE software [49] was used to close gaps from the contigs produced by Velvet. After running IMAGE for 3 iterations using a k-mer size of 55, the number of *pe/ppe* genes assembled increased for all samples, especially in high coverage samples. The majority (range: 78–98 %) of gaps were closed within 3 iterations, which provided a threshold to justify the compromise between runtime and gaps closed in new contigs (fasta format). REAPR software was used to assess the quality of the assemblies, and calculates a quality score per base (http://www.sanger.ac.uk/science/tools/reapr reapr). The final assemblies are available for download (http://pathogenseq.lshtm.ac.uk/ppe). The *pe/ppe* and other genes were called by aligning the assemblies to the well annotated H37Rv genome. The 50,539 SNPs genome-wide were identified using nucmer [50] with H37Rv as the reference genome. To assess the robustness of the aligned sequences and resulting SNPs and analyses, we also mapped samples to a *Mycobacterium africanum* lineage reference (GCA_000253355.1), but observed no major differences from those using H37Rv (lineage 4). Phylogenetic data (alignments, phylogenetic trees) are deposited in Dryad (http://datadryad.org/).

The alignments of the genotypes for the 50,539 SNPs formed the basis of the majority of population genetic analyses, except where stated otherwise. SNP locations at which more than 10 % of the genotypes were missing were excluded from analyses. Other missing data was kept in the multiple alignments and was processed according to the default settings of the analysis software applied. Indels were identified by nucmer but were not analysed in this study. Regions where multiple contigs overlapped or where no contigs mapped to were annotated as missing data. FastTree [51] software employing the generalised time-reversible model was used to produce the final phylogenetic trees. The trees included the ancient *M. canettii* strain (NC_019950.1). The F_{ST} measure was calculated for each SNP to identify markers with complete between-lineage allele differentiation (F_{ST} >0.99). Similarly, the ancestral reconstructed sequence for the lineage-defining node in the phylogenetic tree was compared with its closest ancestral node, and the SNP differences derived. Nucleotide diversity (π) and the number of segregating sites were calculated using variscan software applied to sequence alignments [52]. To test for recombination we used the pairwise homoplasy index (*phi*) statistic calculated in sliding windows, as implemented in Phipack software [27]. The non-synonymous to synonymous ratio was calculated using PAML software [53]. To discover the effect of positive selection on the *pe/ppe* genes compared to all other genes, codeml was used to fit a number of models to the data using a maximum likelihood approach. This is generally thought to be more robust than counting methods. A *dN/dS* (ω) value was calculated per gene across all positions and all branches of the phylogenetic tree. For each gene, we then performed a likelihood ratio test using PAML software to assess evidence of positive selection, which compared two models: (a) variable selective pressure but no positive selection ($0 < \omega < 1$) (M8a) and (b) variable selective pressure with positive selection (M8) ($\omega > 1$). The test statistic has a $\chi 2$ (1 degree of freedom) distribution, and the resulting *p*-value reflects the likelihood of positive selection acting on a gene. To localize the specific polymorphisms under selection we applied the Bayes Empirical Bayes (BEB) method [29]. The proportion of segregating sites under positive selection (S_p/S_s) was calculated using the results from variscan and BEB. Epitopes were predicted using netMHCpan [54] using HLA alleles previously suggested [21].

No ethical approvals were required for this study.

Availability of supporting data

The list of raw sequence data accession numbers for the ENA short read archive, final assemblies and links to the phylogenetic data (alignments, phylogenetic trees) in Dryad can be found in http://pathogenseq.lshtm.ac.uk/ppe.

Additional files

Additional file 1: Table S1. a) The samples used for the assembly (*Malawi [55, 56], Netherlands [57], Pakistan [58], Portugal [59]) and b) the 21 reference strains. **Table S2.** Lineage, sequence coverage and polymorphism. π nucleotide diversity; Lineage 1 Indo-Oceanic; Lineage 2 East-Asian (Beijing); Lineage 3 East-African-Indian; Lineage 4 Euro-American. **Table S3.** Completeness of *pe/ppe* gene assemblies. **Table S4.** List of 87 *pe/ppe* lineage specific-markers. S synonymous, NS non-synonymous, * genes bolded if there are sites under selection using the Bayes Empirical Bayes method; Lineage 1 Indo-Oceanic; Lineage 2 East-Asian (Beijing); Lineage 3 East-African-Indian; Lineage 4 Euro-American. **Table S5.** Genes with more than 10 sites under selective pressure (*dN/dS* (ω) >1). **Table S6.** Epitopes. * identified using netMHCpan, ** epitopes that had sites under positive selection according to the Bayes Empirical Bayes (BEB) method. (DOCX 70 kb)

Additional file 2: Figure S1. Allele frequency spectra for each lineage by synonymous (blue) and non-synonymous (red) mutations. The peaks at intermediate allele frequencies include sub-lineage defining SNPs (Lineage 1 Indo-Oceanic; Lineage 2 East-Asian (Beijing); Lineage 3 East-African-Indian; Lineage 4 Euro-American). (TIF 207 kb)

Additional file 3: Figure S2. Gene-based nucleotide diversity (π) for the 21 reference genomes. All genes with high nucleotide diversity ($\pi > 0.0075$) are labelled. (TIF 148 kb)

Additional file 4: Figure S3. Phylogenetic tree constructed using 50,540 genome-wide SNPs. Clear clustering according to lineage can be seen (Lineage 1 (Indo-Oceanic, green), lineage 2 (East-Asian (Beijing), blue), lineage 3 (East-African-Indian, purple), lineage 4 (Euro-American, red)). Reference genomes are labelled. *M. canetti* is annotated in cyan. (TIF 69 kb)

Additional file 5: Figure S4. Identifying sites leading to differences in tree topologies based on all SNPs (Additional file 4: Figure S3a) and only those from *pe/ppe* genes (Additional file 4: Figure S3b). The Δ Site wise log likelihood score (Δ SSLS) is calculated for each SNP in the *pe/ppe* gene alignments. Negative differences indicate SNP positions favouring the *pe/ppe* tree. SNPs in *pe_pgrs3, ppe57* and *ppe60* produce strong phylogenetic signals supporting the *pe/ppe* tree. (TIF 113 kb)

Additional file 6: Figure S5. Phylogenetic tree created using only SNPs from *pe_pgrs3*. No clear clustering by lineage is observed. However there are two major clades, one consistent with H37Rv (bottom-left). (TIF 126 kb)

Additional file 7: Figure S6. Lineage-specific recombination hotspots. Manhattan plots showing genes that are likely to be recombination hotspots in each lineage (Lineage 1 Indo-Oceanic; Lineage 2 East-Asian (Beijing); Lineage 3 East-African-Indian; Lineage 4 Euro-American). The (−log10) p-value for the *phi* statistic is plotted against genome position. All genes with p-values < 0.05 are labelled. (TIF 147 kb)

Additional file 8: Figure S7. Evidence of recombination at a gene level in the 21 reference genomes. A Manhattan plot showing genes that are likely to be recombination hotspots. The (−log10) p-value for the *phi* statistic is plotted against genome position. Genes with p-values less than 0.05 are shown. (TIF 120 kb)

Additional file 9: Figure S8. Selection *dN/dS* values for each gene within Clusters of Orthologous Groups (COG*) categories. *ppe/N = *pe/ppe* genes annotated as COG category N, * COG categories: **A** RNA processing and modification, **B** Chromatin Structure and dynamics, **C** Energy production and conversion, **D** Cell cycle control and mitosis, **E** Amino Acid metabolism and transport, **F** Nucleotide metabolism and transport, **G** Carbohydrate metabolism and transport, **H** Coenzyme metabolism, **I** Lipid metabolism, **J** Translation, **K** Transcription, **L** Replication and repair, **M** Cell wall/membrane/envelope biogenesis, **N** Cell motility, **O** Post-translational modification, protein turnover, chaperone functions, **P** Inorganic ion transport and metabolism, **Q** Secondary Structure, **T** Signal Transduction, **U** Intracellular trafficking and secretion, **Y** Nuclear structure, **Z** Cytoskeleton, **R** General Functional Prediction only, **S** Function Unknown. (TIF 124 kb)

Additional file 10: Figure S9. Non-neutral evolution for genes within Clusters of Orthologous Groups (COG*) categories. Boxplots are constructed using (-log10) p-values of non-neutral evolution for each gene. *ppe/N = *pe/ppe* genes annotated as COG category N, * COG categories: **A** RNA processing and modification, **B** Chromatin Structure and dynamics, **C** Energy production and conversion, **D** Cell cycle control and mitosis, **E** Amino Acid metabolism and transport, **F** Nucleotide metabolism and transport, **G** Carbohydrate metabolism and transport, **H** Coenzyme metabolism, **I** Lipid metabolism, **J** Translation, **K** Transcription, **L** Replication and repair, **M** Cell wall/membrane/envelope biogenesis, **N** Cell motility, **O** Post-translational modification, protein turnover, chaperone functions, **P** Inorganic ion transport and metabolism, **Q** Secondary Structure, **T** Signal Transduction, **U** Intracellular trafficking and secretion, **Y** Nuclear structure, **Z** Cytoskeleton, **R** General Functional Prediction only, **S** Function Unknown. (TIF 127 kb)

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JEP, FC, RM, and TGC conceived the project. IB, RMA, RW, SLS, NCGvP, JRG, ACC, AA, TBB, SC, KD, LG, RH, ZH, AM, DM, SP, JP, IP, PS, EdOS, EMS, PDvH, MV, and AP contributed isolates, sequencing resources and/or reagents. JEP, FC, SC, MLH and TGC contributed towards data analysis. JEP, RM and TGC wrote the first version of the manuscript, with all other authors contributing to the final version. All authors have read and approve of the final version of the manuscript.

Acknowledgements

JP is supported by a Biotechnology and Biological Sciences Research Council UK PhD studentship. FC was the recipient of a Bloomsbury Research Fund PhD studentship. SLS receives funding from the South African Research Chairs Initiative of the Department of Science and Technology and National Research Foundation (UID 86539). AP is supported by the KAUST faculty baseline research fund (KAUST-BRF). TGC receives funding from the Medical Research Council UK (grant numbers MR/K000551/1, MR/M01360X/1, MR/N010469/1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author details

[1]Department of Pathogen Molecular Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, WC1E 7HT London, UK. [2]KIT Biomedical Research, Royal Tropical Institute, Amsterdam, Netherlands. [3]Department of Science and Technology and National Research Foundation Centre of Excellence for Biomedical Tuberculosis Research, and Medical Research Council Centre for Molecular and Cellular Biology, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, South Africa. [4]Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, WC1E 7HT London, UK. [5]Karonga Prevention Study, Lilongwe, Malawi. [6]National Mycobacterium Reference Laboratory, Porto, Portugal. [7]Centro de Pesquisas Goncalo Moniz, Fundacao Oswaldo Cruz Bahia R, Salvador, Bahia, Brazil. [8]Department of Medicine, Lung Infection and Immunity Unit, Division of Pulmonology & UCT Lung Institute, University of Cape Town, Cape Town, Western Cape, South Africa. [9]Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Cape Town, Western Cape, South Africa. [10]Laboratorio de Enfermedades Infecciosas, Laboratorios de Investigación y Desarrollo, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Lima, Peru. [11]Department of Pathology and Laboratory Medicine, The Aga Khan University, Stadium Road, Karachi, Pakistan. [12]National Center of Infectious and Parasitic Diseases, 1504 Sofia, Bulgaria. [13]Universidade de Lisboa, Lisbon, Portugal. [14]Grupo de Micobactérias, Unidade de Microbiologia Médica, Global Health and Tropical Medicine (GHTM), Instituto de Higiene e Medicina Tropical, Universidade NOVA de Lisboa (IHMT/UNL), Lisbon, Portugal. [15]Biological and Environmental Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia.

Received: 12 November 2015 Accepted: 12 February 2016
Published online: 29 February 2016

References

1. World Health Organization. Global Tuberculosis Report 2014. 2014.
2. Wilkie MEM, McShane H. TB vaccine development: where are we and why is it so difficult? Thorax. 2015;70:299–301.
3. Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, Portugal I, Pain A, Martin N, Clark TG. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. Nat Commun. 2014;5:4812.
4. Gagneux S, Small PM. Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. Lancet Infect Dis. 2007;7:328–37.
5. de Jong BC, Hill PC, Aiken A, Awine T, Antonio M, Adetifa IM, Jackson-Sillah DJ, Fox A, Deriemer K, Gagneux S, Borgdorff MW, McAdam KPWJ, Corrah T, Small PM, Adegbola RA. Progression to active tuberculosis, but not transmission, varies by Mycobacterium tuberculosis lineage in The Gambia. J Infect Dis. 2008; 198:1037–43.
6. Caws M, Thwaites G, Dunstan S, Hawn TR, Lan NTN, Thuong NTT, Stepniewska K, Huyen MNT, Bang ND, Loc TH, Gagneux S, van Soolingen D, Kremer K, van der Sande M, Small P, Anh PTH, Chinh NT, Quy HT, Duyen NTH, Tho DQ, Hieu NT, Torok E, Hien TT, Dung NH, Nhu NTQ, Duy PM, van Vinh Chau N, Farrar J. The influence of host and bacterial genotype on the development of disseminated disease with Mycobacterium tuberculosis. PLoS Pathog. 2008;4:e1000034.
7. Ordway DJ, Shang S, Henao-Tamayo M, Obregon-Henao A, Nold L, Caraway M, Shanley CA, Basaraba RJ, Duncan CG, Orme IM. Mycobacterium bovis BCG-mediated protection against W-Beijing strains of Mycobacterium tuberculosis is diminished concomitant with the emergence of regulatory T cells. Clin Vaccine Immunol. 2011;18:1527–35.
8. Niemann S, Supply P. Diversity and evolution of Mycobacterium tuberculosis: moving to whole-genome-based approaches. Cold Spring Harb Perspect Med. 2014;4:a021188.
9. Fishbein S, van Wyk N, Warren RM, Sampson SL. Phylogeny to function: PE/PPE protein evolution and impact on Mycobacterium tuberculosis pathogenicity. Mol Microbiol. 2015.
10. Akhter Y, Ehebauer MT, Mukhopadhyay S, Hasnain SE. The PE/PPE multigene family codes for virulence factors and is a possible source of mycobacterial antigenic variation: perhaps more? Biochimie. 2012;94:110–6.
11. van Pittius NC G, Sampson SL, Lee H, Kim Y, van Helden PD, Warren RM. Evolution and expansion of the Mycobacterium tuberculosis PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions. BMC Evol Biol. 2006;6:95.
12. Bottai D, Di Luca M, Majlessi L, Frigui W, Simeone R, Sayes F, Bitter W, Brennan MJ, Leclerc C, Batoni G, Campa M, Brosch R, Esin S. Disruption of the ESX-5 system of Mycobacterium tuberculosis causes loss of PPE protein secretion, reduction of cell wall integrity and strong attenuation. Mol Microbiol. 2012;83:1195–209.
13. Majlessi L, Prados-Rosales R, Casadevall A, Brosch R. Release of mycobacterial antigens. Immunol Rev. 2015;264:25–45.
14. Delogu G, Cole ST, Brosch R. The PE and PPE Protein Families of Mycobacterium Tuberculosis. In: Handbook of Tuberculosis: Molecular Biology and Biochemistryitle. 2008. p. 131–50.
15. Mohareer K, Tundup S, Hasnain SE. Transcriptional regulation of Mycobacterium tuberculosis PE/PPE genes: a molecular switch to virulence? J Mol Microbiol Biotechnol. 2011;21:97–109.
16. Wang H, Dong D, Tang S, Chen X, Gao Q. PPE38 of Mycobacterium marinum triggers the cross-talk of multiple pathways involved in the host response, as revealed by subcellular quantitative proteomics. J Proteome Res. 2013;12: 2055–66.
17. Singh KK, Zhang X, Patibandla AS, Chien P, Laal S. Antigens of Mycobacterium tuberculosis expressed during preclinical tuberculosis: serological immunodominance of proteins with repetitive amino acid sequences. Infect Immun. 2001;69:4185–91.
18. Galagan JE. Genomic insights into tuberculosis. Nat Rev Genet. 2014;15: 307–20.
19. Gutacker MM, Mathema B, Soini H, Shashkina E, Kreiswirth BN, Graviss EA, Musser JM. Single-nucleotide polymorphism-based population genetic analysis of Mycobacterium tuberculosis strains from 4 geographic sites. J Infect Dis. 2006;193:121–8.
20. Musser JM, Amin A, Ramaswamy S. Negligible Genetic Diversity of Mycobacterium tuberculosis Host Immune System Protein Targets: Evidence of Limited Selective Pressure. Genetics. 2000;155:7–16.
21. Copin R, Coscollá M, Seiffert SN, Bothamley G, Sutherland J, Mbayo G, Gagneux S, Ernst JD. Sequence diversity in the pe_pgrs genes of Mycobacterium tuberculosis is independent of human T cell recognition. MBio. 2014;5:e00960–13.
22. Mycobacterium tuberculosis 18b genome. [http://www.ncbi.nlm.nih.gov/nuccore/CP007299.1]
23. Rodríguez JG, Pino C, Tauch A, Murcia MI. Complete Genome Sequence of the Clinical Beijing-Like Strain Mycobacterium tuberculosis 323 Using the PacBio Real-Time Sequencing Platform. Genome Announc. 2015;3.
24. Pepperell CS, Casto AM, Kitchen A, Granka JM, Cornejo OE, Holmes EC, Birren B, Galagan J, Feldman MW. The role of selection in shaping diversity of natural M. tuberculosis populations. PLoS Pathog. 2013;9:e1003543.
25. McEvoy CRE, Cloete R, Müller B, Schürch AC, van Helden PD, Gagneux S, Warren RM, Gey van Pittius NC. Comparative analysis of Mycobacterium tuberculosis pe and ppe genes reveals high sequence variation and an apparent absence of selective constraints. PLoS One. 2012;7:e30593.
26. Liu X, Gutacker MM, Musser JM, Fu Y-X. Evidence for recombination in Mycobacterium tuberculosis. J Bacteriol. 2006;188:8169–77.
27. Bruen TC, Philippe H, Bryant D. A simple and robust statistical test for detecting the presence of recombination. Genetics. 2006;172:2665–81.
28. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 2000;28:33–6.
29. Yang Z, Wong WSW, Nielsen R. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol. 2005;22:1107–18.
30. Graña M, Bellinzoni M, Bellalou J, Haouz A, Miras I, Buschiazzo A, Winter N, Alzari PM. Crystal structure of Mycobacterium tuberculosis LppA, a lipoprotein confined to pathogenic mycobacteria. Proteins. 2010;78:769–72.
31. 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. Variable host-pathogen compatibility in Mycobacterium tuberculosis. Proc Natl Acad Sci U S A. 2006;103:2869–73.
32. Tsolaki AG, Hirsh AE, DeRiemer K, Enciso JA, Wong MZ, Hannan M, Goguet de la Salmoniere Y-OL, Aman K, Kato-Maeda M, Small PM. Functional and evolutionary genomics of Mycobacterium tuberculosis: insights from genomic deletions in 100 strains. Proc Natl Acad Sci U S A. 2004;101: 4865–70.
33. Målen H, Berven FS, Fladmark KE, Wiker HG. Comprehensive analysis of exported proteins from Mycobacterium tuberculosis H37Rv. Proteomics. 2007;7:1702–18.
34. Sampson SL. Mycobacterial PE/PPE proteins at the host-pathogen interface. Clin Dev Immunol. 2011;2011:497203.
35. Karboul A, van Pittius NC G, Namouchi A, Vincent V, Sola C, Rastogi N, Suffys P, Fabre M, Cataldi A, Huard RC, Kurepina N, Kreiswirth B, Ho JL, Gutierrez MC, Mardassi H. Insights into the evolutionary history of tubercle bacilli as disclosed by genetic rearrangements within a PE_PGRS duplicated gene pair. BMC Evol Biol. 2006;6:107.
36. Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, Feil EJ. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol. 2006;239:226–35.
37. Chaitra MG, Shaila MS, Nayak R. Evaluation of T-cell responses to peptides with MHC class I-binding motifs derived from PE_PGRS 33 protein of Mycobacterium tuberculosis. J Med Microbiol. 2007;56(Pt 4):466–74.
38. Comas I, Chakravartti J, Small PM, Galagan J, Niemann S, Kremer K, Ernst JD, Gagneux S. Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. Nat Genet. 2010;42:498–503.
39. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, Warren RM, Streicher EM, Calver A, Sloutsky A, Kaur D, Posey JE, Plikaytis B, Oggioni MR, Gardy JL, Johnston JC, Rodrigues M, Tang PKC, Kato-Maeda M, Borowsky ML, Muddukrishna B, Kreiswirth BN, Kurepina N, Galagan J, Gagneux S, Birren B, Rubin EJ, Lander ES, Sabeti PC, Murray M. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nat Genet. 2013;45:1183–9.
40. Mehaffy C, Guthrie JL, Alexander DC, Stuart R, Rea E, Jamieson FB. Marked microevolution of a unique Mycobacterium tuberculosis strain in 17 years of ongoing transmission in a high risk population. PLoS One. 2014;9:e112928.
41. Coll F, Preston M, Guerra-Assunção JA, Hill-Cawthorn G, Harris D, Perdigão J, Viveiros M, Portugal I, Drobniewski F, Gagneux S, Glynn JR, Pain A, Parkhill J,

McNerney R, Martin N, Clark TG. PolyTB: a genomic variation map for Mycobacterium tuberculosis. Tuberculosis (Edinb). 2014;94:346–54.

42. Coll F, McNerney R, Preston M, Guerra-Assunção JA, Warry A, Hill-Cawthorn G, Mallard K, Nair M, Miranda A, Alves A, Perdigão J, Viveiros M, Portugal I, Hasan Z, Hasan R, Glynn JR, Martin N, Pain A, Clark TG. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med. 2015, In Press.
43. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013;29:2669–77.
44. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 2012;22:549–56.
45. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18:821–9.
46. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin A V, Sirotkin A V, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455–77.
47. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–5.
48. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
49. Tsai IJ, Otto TD, Berriman M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biol. 2010;11:R41.
50. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. Genome Biol. 2004;5:R12.
51. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009;26:1641–50.
52. Vilella AJ, Blanco-Garcia A, Hutter S, Rozas J. VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. Bioinformatics. 2005;21:2791–3.
53. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24:1586–91.
54. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, Røder G, Peters B, Sette A, Lund O, Buus S. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. PLoS One. 2007;2:e796.
55. Guerra-Assunção JA, Houben RMGJ, Crampin AC, Mzembe T, Mallard K, Coll F, et al. Recurrence due to Relapse or Reinfection With Mycobacterium tuberculosis: A Whole-Genome Sequencing Approach in a Large, Population-Based Cohort With a High HIV Infection Prevalence and Active Follow-up. J Infect Dis. 2014. doi:10.1093/infdis/jiu574.
56. Guerra-Assunção J, Crampin A, Houben R, Mzembe T, Mallard K, Coll F, et al. Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. Elife. 2015;4. doi:10.7554/eLife.05166.
57. Bergval I, Coll F, Schuitema A, de Ronde H, Mallard K, Pain A, et al. A proportion of mutations fixed in the genomes of in vitro selected isogenic drug-resistant Mycobacterium tuberculosis mutants can be detected as minority variants in the parent culture. FEMS Microbiol Lett. 2015;362:1–7. doi:10.1093/femsle/fnu037.
58. Hasan Z, Ali A, McNerney R, Mallard K, Hill-Cawthorne G, Coll F, et al. Whole genome sequencing-based characterization of extensively drug resistant (XDR) strains of Mycobacterium tuberculosis from Pakistan. Int J Mycobacteriology Elsevier. 2015;4:11–2. doi:10.1016/j.ijmyco.2014.10.050.
59. Perdigão J, Silva H, Machado D, Macedo R, Maltez F, Silva C, et al. Unraveling Mycobacterium tuberculosis genomic diversity and evolution in Lisbon, Portugal, a highly drug resistant setting. BMC Genomics. 2014;15:991. doi:10.1186/1471-2164-15-991.

Supplementary File 1: S1 Table
a) The samples used for the assembly (*Malawi [55, 56], Netherlands [57], Pakistan [58], Portugal [59]) and b) the 21 reference strains.
a)

				Lineage 3	Lineage 4
Study	No.	Lineage 1	Lineage 2	East-African-	Euro-
location*	samples	Indo-Oceanic	East-Asian	Indian	American
Brazil	42	-	-	-	42
Bulgaria	2	-	-	-	2
China	6	-	-	5	1
Malawi	257	38	8	28	183
Netherlands	10	-	-	-	10
Pakistan	31	4	4	19	4
Peru	65	-	5	-	60
Portugal	78	-	5	-	73
South Africa	27	-	16	1	10
Total	**518**	**42**	**38**	**53**	**385**

*Malawi [56, 57], Netherlands [58], Pakistan [59], Portugal [60]

b)

Strain	Assembly Accession	Lineage
CDC1551	GCA_000008585.1	Lineage4
CTRI_2	GCA_000224435.1	Lineage4
F11	GCA_000016925.1	Lineage4
7199_99	GCA_000331445.1	Lineage4
H37Ra	GCA_000016145.1	Lineage4
KZN_1435	GCA_000023625.1	Lineage4
KZN_4207	GCA_000154585.2	Lineage4
KZN_605	GCA_000154605.2	Lineage4
RGTB327	GCA_000277085.1	Lineage4
RGTB423	GCA_000277105.1	Lineage1
Beijing_NITR203	GCA_000364825.1	Lineage2
Erdman_ATCC_35801	GCA_000350205.1	Lineage4
Haarlem	GCA_000153685.2	Lineage4
UT205	GCA_000304555.1	Lineage4
W_148	GCA_000193185.1	Lineage2
CAS_NITR204	GCA_000389925.1	Lineage3
CCDC5079	GCA_000270345.1	Lineage2
CCDC5180	GCA_000270365.1	Lineage2
M.bovis_Pasteur_1173p2	GCA_000009445.1	Bovis
M323	Genbank CP010873.1	Lineage 2
18b	Genbank CP007299.1	Lineage 2

Supplementary file 1: S2 Table
Lineage, sequence coverage and polymorphism. *π* nucleotide diversity; Lineage 1 Indo-Oceanic; Lineage 2 East-Asian (Beijing); Lineage 3 East-African-Indian; Lineage 4 Euro-American.

Lineage	*n* (%)	Median Coverage across genome	Median Coverage across *pe/ppe* genes	Median *π* across genome	Median *π* across *pe/ppe* genes	No. Lineage specific *pe/ppe* SNPs
1	42 (8.1)	187.7	127.7	0.00009	0.00017	36
2	38 (7.3)	319.7	151.1	0.00002	0.00007	15
3	53 (10.2)	329.5	174.0	0.00004	0.00007	28
4	385 (74.3)	268.3	150.9	0.00007	0.00016	8
Overall	518	283.5	155.4	0.00014	0.00027	87

Supplementary file 1: S3 Table
Completeness of *pe/ppe* gene assemblies.

Locus	Gene	Length	Total length of gaps (prop. of gene length)	Proportion of samples fully assembled	No. SNPs	Non-synonymous SNPs
Rv0109	*pe_pgrs1*	1490	0 (0)	0.99	23	13
Rv0124	*pe_pgrs2*	1463	0 (0)	0.93	27	18
Rv0151c	*pe1*	1766	0 (0)	0.99	32	20
Rv0152c	*pe2*	1577	0 (0)	1	21	15
Rv0159c	*pe3*	1406	0 (0)	1	18	13
Rv0160c	*pe4*	1508	0 (0)	1	16	10
Rv0278c	*pe_pgrs3*	2873	0 (0)	0.78	281	135
Rv0279c	*pe_pgrs4*	2513	241 (0.1)	0.25	110	52
Rv0285	*pe5*	308	0 (0)	1	4	2
Rv0297	*pe_pgrs5*	1775	0 (0)	0.98	23	16
Rv0335c	*pe6*	515	0 (0)	1	28	17
Rv0532	*pe_pgrs6*	1784	0 (0)	0.95	69	46
Rv0578c	*pe_pgrs7*	3920	0 (0)	0.75	120	55
Rv0742	*pe_pgrs8*	527	0 (0)	0.99	3	2
Rv0746	*pe_pgrs9*	2351	23 (0.01)	0.44	68	41
Rv0747	*pe_pgrs10*	2405	0 (0)	0.56	188	100
Rv0754	*pe_pgrs11*	1754	0 (0)	1	13	8
Rv0832	*pe_pgrs12*	413	0 (0)	1	2	2
Rv0833	*pe_pgrs13*	2249	0 (0)	0.77	63	42
Rv0834c	*pe_pgrs14*	2648	0 (0)	0.92	62	26
Rv0872c	*pe_pgrs15*	1820	0 (0)	1	16	9
Rv0916c	*pe7*	299	0 (0)	1	3	3
Rv0977	*pe_pgrs16*	2771	0 (0)	0.82	136	103
Rv0978c	*pe_pgrs17*	995	0 (0)	0.51	33	19
Rv0980c	*pe_pgrs18*	1373	318 (0.23)	0.14	48	26
Rv1040c	*pe8*	827	0 (0)	1	4	3
Rv1067c	*pe_pgrs19*	2003	305.5 (0.15)	0.12	81	40
Rv1068c	*pe_pgrs20*	1391	207 (0.15)	0.2	5	5
Rv1087	*pe_pgrs21*	2303	0 (0)	0.58	77	48
Rv1088	*pe9*	434	0 (0)	1	3	2
Rv1089	*pe10*	362	0 (0)	1	5	4
Rv1091	*pe_pgrs22*	2561	197 (0.08)	0.13	55	28
Rv1172c	*pe12*	926	0 (0)	1	6	3
Rv1195	*pe13*	299	0 (0)	1	12	9
Rv1214c	*pe14*	332	0 (0)	1	4	2
Rv1243c	*pe_pgrs23*	1688	0 (0)	0.91	11	8

Rv1325c	*pe_pgrs24*	1811	0 (0)	0.88	54	23
Rv1386	*pe15*	308	0 (0)	1	3	2
Rv1396c	*pe_pgrs25*	1730	0 (0)	0.96	36	20
Rv1430	*pe16*	1586	0 (0)	1	12	10
Rv1441c	*pe_pgrs26*	1475	0 (0)	0.85	14	10
Rv1450c	*pe_pgrs27*	3989	418 (0.1)	0.3	55	29
Rv1452c	*pe_pgrs28*	2225	22 (0.01)	0.49	51	19
Rv1468c	*pe_pgrs29*	1112	0 (0)	1	16	7
Rv1646	*pe17*	932	0 (0)	1	2	2
Rv1651c	*pe_pgrs30*	3035	0 (0)	0.98	40	20
Rv1768	*pe_pgrs31*	1856	0 (0)	0.97	22	17
Rv1788	*pe18*	299	0 (0)	0.95	15	14
Rv1791	*pe19*	299	0 (0)	0.99	17	13
Rv1803c	*pe_pgrs32*	1919	0 (0)	1	27	17
Rv1806	*pe20*	299	0 (0)	1	3	2
Rv1818c	*pe_pgrs33*	1496	0 (0)	0.98	36	14
Rv1840c	*pe_pgrs34*	1547	0 (0)	0.99	22	13
Rv1983	*pe_pgrs35*	1676	0 (0)	1	14	9
Rv2098c	*pe_pgrs36*	1304	0 (0)	0.99	7	5
Rv2099c	*pe21*	173	0 (0)	1	3	2
Rv2107	*pe22*	296	0 (0)	1	2	1
Rv2126c	*pe_pgrs37*	770	0 (0)	0.99	21	12
Rv2162c	*pe_pgrs38*	1598	0 (0)	0.79	45	16
Rv2328	*pe23*	1148	0 (0)	1	9	7
Rv2340c	*pe_pgrs39*	1241	0 (0)	1	16	9
Rv2371	*pe_pgrs40*	185	0 (0)	1	1	0
Rv2396	*pe_pgrs41*	1085	0 (0)	0.91	26	15
Rv2408	*pe24*	719	0 (0)	1	5	4
Rv2431c	*pe25*	299	0 (0)	1	3	2
Rv2487c	*pe_pgrs42*	2084	0 (0)	0.85	21	10
Rv2490c	*pe_pgrs43*	4982	14 (0)	0.43	103	44
Rv2519	*pe26*	1478	0 (0)	1	19	11
Rv2591	*pe_pgrs44*	1631	0 (0)	0.96	19	12
Rv2615c	*pe_pgrs45*	1385	0 (0)	0.51	27	11
Rv2634c	*pe_pgrs46*	2336	0 (0)	0.97	21	10
Rv2741	*pe_pgrs47*	1577	0 (0)	0.86	56	33
Rv2769c	*pe27*	827	0 (0)	1	13	11
Rv2853	*pe_pgrs48*	1847	0 (0)	0.98	24	17
Rv3018A	*pe27A*	86	0 (0)	0.69	0	0
Rv3022A	*pe29*	314	0 (0)	0.98	0	0
Rv3344c	*pe_pgrs49*	1454	0 (0)	0.78	47	16
Rv3345c	*pe_pgrs50*	4616	125 (0.03)	0.22	207	105
Rv3367	*pe_pgrs51*	1766	0 (0)	1	15	7

Rv3388	*pe_pgrs52*	2195	0 (0)	0.76	48	33
Rv3477	*pe31*	296	0 (0)	1	10	8
Rv3507	*pe_pgrs53*	4145	0 (0)	0.64	133	97
Rv3508	*pe_pgrs54*	5705	2018 (0.35)	0	358	200
Rv3511	*pe_pgrs55*	2144	233 (0.11)	0.2	119	74
Rv3512	*pe_pgrs56*	3239	345 (0.11)	0.04	174	114
Rv3514	*pe_pgrs57*	4469	2651 (0.59)	0	39	29
Rv3590c	*pe_pgrs58*	1754	8 (0)	0.48	49	17
Rv3595c	*pe_pgrs59*	1319	0 (0)	1	22	8
Rv3622c	*pe32*	299	0 (0)	0.99	3	2
Rv3650	*pe33*	284	0 (0)	1	4	3
Rv3652	*pe_pgrs60*	314	0 (0)	1	7	4
Rv3653	*pe_pgrs61*	587	0 (0)	0.99	10	8
Rv3746c	*pe34*	335	0 (0)	1	8	8
Rv3812	*pe_pgrs62*	1514	0 (0)	1	20	14
Rv3872	*pe35*	299	0 (0)	1	3	3
Rv3893c	*pe36*	233	0 (0)	1	1	1
Rv0096	*ppe1*	1391	0 (0)	1	27	19
Rv0256c	*ppe2*	1670	0 (0)	1	17	9
Rv0280	*ppe3*	1610	0 (0)	1	18	12
Rv0286	*ppe4*	1541	0 (0)	1	16	11
Rv0304c	*ppe5*	6614	0 (0)	0.97	65	37
Rv0305c	*ppe6*	2891	0 (0)	1	36	22
Rv0354c	*ppe7*	425	0 (0)	1	4	3
Rv0355c	*ppe8*	9902	12 (0)	0.46	329	189
Rv0388c	*ppe9*	542	0 (0)	1	11	2
Rv0442c	*ppe10*	1463	0 (0)	0.98	14	9
Rv0453	*ppe11*	1556	0 (0)	1	18	12
Rv0755c	*ppe12*	1937	0 (0)	1	20	11
Rv0878c	*ppe13*	1331	0 (0)	1	13	8
Rv0915c	*ppe14*	1271	0 (0)	1	12	8
Rv1039c	*ppe15*	1175	0 (0)	1	9	7
Rv1135c	*ppe16*	1856	0 (0)	1	22	18
Rv1168c	*ppe17*	1040	0 (0)	1	9	7
Rv1196	*ppe18*	1175	0 (0)	0.56	6	2
Rv1361c	*ppe19*	1190	0 (0)	0.7	53	31
Rv1387	*ppe20*	1619	0 (0)	1	16	11
Rv1548c	*ppe21*	2036	0 (0)	0.99	24	19
Rv1705c	*ppe22*	1157	0 (0)	0.99	20	13
Rv1706c	*ppe23*	1184	0 (0)	0.99	10	4
Rv1753c	*ppe24*	3161	282 (0.09)	0	68	35
Rv1787	*ppe25*	1097	375 (0.34)	0.36	6	5
Rv1789	*ppe26*	1181	0 (0)	0.94	13	8

Rv1790	*ppe27*	1052	0 (0)	0.53	12	10
Rv1800	*ppe28*	1967	0 (0)	1	33	27
Rv1801	*ppe29*	1271	0 (0)	0.99	9	4
Rv1802	*ppe30*	1391	0 (0)	1	25	17
Rv1807	*ppe31*	1199	0 (0)	1	12	6
Rv1808	*ppe32*	1229	0 (0)	1	13	5
Rv1809	*ppe33*	1406	0 (0)	1	14	9
Rv1917c	*ppe34*	4379	348 (0.08)	0	132	63
Rv1918c	*ppe35*	2963	0 (0)	0.98	54	34
Rv2108	*ppe36*	731	0 (0)	1	10	7
Rv2123	*ppe37*	1421	0 (0)	0.99	20	14
Rv2352c	*ppe38*	1175	0 (0)	0.92	37	15
Rv2353c	*ppe39*	1064	0 (0)	0.56	73	40
Rv2356c	*ppe40*	1847	0 (0)	0.95	23	11
Rv2430c	*ppe41*	584	0 (0)	0.92	6	3
Rv2608	*ppe42*	1742	0 (0)	1	11	5
Rv2768c	*ppe43*	1184	0 (0)	1	17	12
Rv2770c	*ppe44*	1148	0 (0)	1	14	10
Rv2892c	*ppe45*	1226	0 (0)	0.99	10	7
Rv3018c	*ppe46*	1304	151.5 (0.12)	0.45	22	12
Rv3021c	*ppe47*	1076	223.5 (0.21)	0.09	4	1
Rv3022c	*ppe48*	242	133.5 (0.55)	0.36	1	0
Rv3125c	*ppe49*	1175	0 (0)	0.98	26	17
Rv3135	*ppe50*	398	0 (0)	0.68	0	0
Rv3136	*ppe51*	1142	0 (0)	1	17	10
Rv3144c	*ppe52*	1229	0 (0)	1	10	6
Rv3159c	*ppe53*	1772	0 (0)	0.99	24	13
Rv3343c	*ppe54*	7571	543.5 (0.07)	0.06	163	80
Rv3347c	*ppe55*	9473	0 (0)	0.55	0	0
Rv3350c	*ppe56*	11150	0 (0)	0.54	0	0
Rv3425	*ppe57*	530	2 (0)	0.46	37	35
Rv3426	*ppe58*	698	697 (1)	0.49	0	0
Rv3429	*ppe59*	536	0 (0)	0.89	86	73
Rv3478	*ppe60*	1181	0 (0)	0.92	155	110
Rv3532	*ppe61*	1220	0 (0)	1	11	9
Rv3533c	*ppe62*	1748	0 (0)	0.99	16	6
Rv3539	*ppe63*	1439	0 (0)	1	13	9
Rv3558	*ppe64*	1658	0 (0)	1	17	14
Rv3621c	*ppe65*	1241	0 (0)	1	19	14
Rv3738c	*ppe66*	947	0 (0)	0.9	0	0
Rv3739c	*ppe67*	233	0 (0)	0.9	2	1
Rv3873	*ppe68*	1106	0 (0)	1	13	7
Rv3892c	*ppe69*	1199	0 (0)	1	12	6

Rv1169c	*pe11*	302	0 (0)	1	2	2
Rv3020c	*pe28*	293	0 (0)	0.73	0	0
Rv3097c	*pe_pgrs63*	1313	0 (0)	1	12	6

Supplementary file 1: S4 Table

List of 87 *pe/ppe* lineage specific-markers. S synonymous, NS non-synonymous, * genes bolded if there are sites under selection using the Bayes Empirical Bayes method; Lineage 1 Indo-Oceanic; Lineage 2 East-Asian (Beijing); Lineage 3 East-African-Indian; Lineage 4 Euro-American.

Position	Mutation	Locus Tag	Gene	NS/S	Lineage
132646	G/T	*Rv0109*	*pe_pgrs1*	NS	1
189948	C/G	*Rv0160c*	*pe4*	S	1
308312	G/A	*Rv0256c*	*ppe2*	S	1
339508	C/T	*Rv0280*	*ppe3*	S	1
362007	G/A	*Rv0297*	*pe_pgrs5*	NS	1
368948	T/C	*Rv0304c*	*ppe5*	NS	1
372149	G/A	*Rv0304c*	*ppe5*	NS	1
426768	C/T	*Rv0355c*	*ppe8*	NS	1
434327	A/G	*Rv0355c*	*ppe8*	NS	1
673066	C/G	*Rv0578c*	*pe_ pgrs7*	S	1
673344	A/T	*Rv0578c*	*pe_ pgrs7*	S	1
846996	G/A	*Rv0754*	*pe_ pgrs11*	NS	1
928483	C/T	*Rv0834c*	*pe_pgrs14*	NS	1
977196	G/A	*Rv0878c*	*ppe13*	S	1
1188917	G/A	*Rv1067c*	*pe_pgrs19*	NS	1
1656178	C/T	*Rv1468c*	*pe_pgrs29*	NS	1
1863660	C/T	*Rv1651c*	*pe_pgrs30*	NS	1
2045849	C/T	*Rv1803c*	*pe_pgrs32*	NS	1
2165256	T/G	*Rv1917c*	*ppe34*	NS	1
2423785	C/T	*Rv2162c*	*pe_pgrs38*	NS	1
2803867	G/C	*Rv2490c*	*pe_ pgrs43*	S	1
2961099	G/A	*Rv2634c*	*pe_ pgrs46*	NS	1
3053973	C/T	*Rv2741*	*pe_ pgrs47*	S	1
3080282	C/A	*Rv2770c*	*ppe44*	NS	1
3929996	G/T	*Rv3507*	*pe_*pgrs*53*	NS	1
3936696	A/G	*Rv3508*	*pe_pgrs54*	NS	1
3942239	C/A	*Rv3512*	*pe_pgrs56*	S	1
3944807	T/C	*Rv3512*	*pe_pgrs56*	S	1
3970112	C/T	*Rv3532*	*ppe61*	NS	1
3979151	T/A	*Rv3539*	*ppe63*	NS	1
3998895	G/A	*Rv3558*	*ppe64*	NS	1
4061113	G/T	*Rv3621c*	*ppe65*	S	1
4093719	G/A	*Rv3652*	*pe_pgrs60*	NS	1
4277032	G/C	*Rv3812*	*pe_pgrs62*	NS	1
4351759	G/C	*Rv3873*	*ppe68*	NS	1
4375318	G/A	*Rv3892c*	*ppe69*	NS	1

424981	G/A	*Rv0355c*	*ppe8*	S	2
1212432	C/A	*Rv1087*	*pe_pgrs21*	S	2
1217065	C/A	*Rv1091*	*pe_pgrs22*	S	2
1217157	A/C	*Rv1091*	*pe_pgrs22*	NS	2
1218658	G/C	*Rv1091*	*pe_pgrs22*	S	2
1299305	G/A	*Rv1168c*	*ppe17*	NS	2
1357308	T/G	*Rv1214c*	*pe14*	S	2
1606673	G/T	*Rv1430*	*pe16*	S	2
2601760	G/A	*Rv2328*	*pe23*	NS	2
2706663	G/T	*Rv2408*	*pe24*	NS	2
2922846	C/T	*Rv2591*	*pe_pgrs44*	S	2
2922848	A/T	*Rv2591*	*pe_pgrs44*	NS	2
3895585	C/T	*Rv3478*	*ppe60*	NS	2
4032218	G/A	*Rv3590c*	*pe_pgrs58*	NS	2
4032625	G/T	*Rv3590c*	*pe_pgrs58*	S	2
178205	C/G	*Rv0151c*	*pe1*	S	3
178453	C/G	*Rv0151c*	*pe1*	S	3
188317	A/G	*Rv0159c*	*pe3*	NS	3
189850	A/G	*Rv0160c*	*pe4*	NS	3
308661	A/G	*Rv0256c*	*ppe2*	NS	3
350088	C/A	*Rv0286*	*ppe4*	NS	3
367718	G/T	*Rv0304c*	*ppe5*	NS	3
369886	C/G	*Rv0304c*	*ppe5*	S	3
428921	G/A	*Rv0355c*	*ppe8*	NS	3
432459	C/T	*Rv0355c*	*ppe8*	NS	3
531775	C/G	*Rv0442c*	*ppe10*	S	3
623163	C/T	*Rv0532*	*pe_pgrs6*	NS	3
674702	A/T	*Rv0578c*	*pe_pgrs7*	S	3
840847	C/T	*Rv0747*	*pe_pgrs10*	S	3
1488428	C/T	*Rv1325c*	*pe_pgrs24*	NS	3
1489142	C/T	*Rv1325c*	*pe_pgrs24*	NS	3
1856617	C/T	*Rv1646*	*pe17*	NS	3
1863584	G/T	*Rv1651c*	*pe_pgrs30*	NS	3
2051345	G/A	*Rv1809*	*ppe33*	NS	3
2382289	G/T	*Rv2123*	*ppe37*	NS	3
2836773	C/T	*Rv2519*	*pe26*	NS	3
2943675	G/A	*Rv2615c*	*pe_pgrs45*	S	3
2960592	C/T	*Rv2634c*	*pe_pgrs46*	NS	3
3738364	G/A	*Rv3344c*	*ppe52*	NS	3
3738364	G/A	*Rv3345c*	*pe_pgrs50*	S	3
3740181	T/C	*Rv3345c*	*pe_pgrs50*	NS	3
3741240	C/T	*Rv3345c*	*pe_pgrs50*	NS	3
4375452	G/A	*Rv3892c*	*ppe69*	NS	3

428698	T/C	*Rv0355c*	*ppe8*	S	4
1618978	C/T	*Rv1441c*	*pe_pgrs26*	NS	4
1931718	C/G	*Rv1705c*	*ppe22*	S	4
2050822	C/G	*Rv1808*	*ppe32*	NS	4
2167926	G/A	*Rv1918c*	*ppe35*	NS	4
3079877	G/A	*Rv2770c*	*ppe44*	NS	4
3466919	G/C	*Rv3097c*	*lipY*	S	4
3510120	G/T	*Rv3144c*	*ppe52*	NS	4

Supplementary file 1: S5 Table
Genes with more than 10 sites under selective pressure (*dN/dS (*ω) >1).

Name	Locus	Length	Function	No. sites
pe_pgrs3	*Rv0278c*	2873	*pe/ppe*	49
pe_pgrs54	*Rv3508*	5705	*pe/ppe*	39
Rv0668	*Rv0668*	3950	information pathways	34
pe_pgrs57	*Rv3514*	4469	*pe/ppe*	33
ppe54	*Rv3343c*	7571	*pe/ppe*	32
pe_pgrs56	*Rv3512*	3239	*pe/ppe*	29
ppe55	*Rv3347c*	9473	*pe/ppe*	29
ppe56	*Rv3350c*	11150	*pe/ppe*	26
pks12	*Rv2048c*	12455	lipid metabolism	25
pe_pgrs28	*Rv1452c*	2225	*pe/ppe*	23
Rv2850c	*Rv2850c*	1889	metabolism & respiration	21
Rv0075	*Rv0075*	1172	metabolism & respiration	20
lppA	*Rv2543*	659	cell wall & cell processes	20
lppB	*Rv2544*	662	cell wall & cell processes	19
pe_pgrs50	*Rv3345c*	4616	*pe/ppe*	18
ppe57	*Rv3425*	530	*pe/ppe*	18
Rv1453	*Rv1453*	1265	regulatory proteins	18
ppsA	*Rv2931*	5630	lipid metabolism	18
Rv1722	*Rv1722*	1484	lipid metabolism	17
ctpJ	*Rv3743c*	1982	cell wall & cell processes	17
pe_pgrs17	*Rv0978c*	995	*pe/ppe*	16
pe_pgrs18	*Rv0980c*	1373	*pe/ppe*	16
fadE1	*Rv0131c*	1343	lipid metabolism	16
Rv1729c	*Rv1729c*	938	lipid metabolism	16
pe_pgrs19	*Rv1067c*	2003	*pe/ppe*	15
pe_pgrs4	*Rv0279c*	2513	*pe/ppe*	15
pe_pgrs16	*Rv0977*	2771	*pe/ppe*	14
Rv2978c	*Rv2978c*	1379	insertion sequences & phages	14
pe_pgrs21	*Rv1087*	2303	*pe/ppe*	13
pe_pgrs9	*Rv0746*	2351	*pe/ppe*	13
ppe8	*Rv0355c*	9902	*pe/ppe*	13
Rv0080	*Rv0080*	458	NA	13
treY	*Rv1563c*	2297	virulence, detoxification & adaptation	13
Rv2827c	*Rv2827c*	887	NA	13
Rv2082	*Rv2082*	2165	NA	12
pe_pgrs10	*Rv0747*	2405	*pe/ppe*	11
ppe10	*Rv0442c*	1463	*pe/ppe*	11
Rv0893c	*Rv0893c*	977	lipid metabolism	11
Rv1254	*Rv1254*	1151	metabolism & respiration	11

Rv1776c	*Rv1776c*	560	regulatory proteins	11
acrA1	*Rv3391*	1952	lipid metabolism	11

Supplementary file 1: S6 Table

Epitopes. * identified using netMHCpan, ** epitopes that had sites under positive selection according to the Bayes Empirical Bayes (BEB) method.

Gene	No. epitopes found*	No. (%) sites disturbed**
pe_pgrs49	**2**	**2 (100)**
ppe59	**55**	**45 (81.8)**
ppe60	**95**	**61 (64.2)**
pe_pgrs60	**15**	**5 (33.3)**
pe18	**16**	**5 (31.3)**
pe_pgrs26	**43**	**12 (27.9)**
ppe57	**38**	**10 (26.3)**
pe6	**40**	**9 (22.5)**
ppe65	**85**	**19 (22.4)**
ppe27	**105**	**22 (21)**
pe_pgrs12	**35**	**7 (20)**
pe25	17	3 (17.6)
pe_pgrs7	26	4 (15.4)
pe_pgrs20	29	4 (13.8)
ppe54	373	44 (11.8)
ppe46	125	14 (11.2)
ppe19	101	11 (10.9)
ppe47	83	9 (10.8)
ppe22	95	10 (10.5)
pe_pgrs10	40	4 (10)
ppe52	61	6 (9.8)
pe_pgrs13	11	1 (9.1)
pe_pgrs3	67	6 (9)
ppe13	84	7 (8.3)
pe3	123	10 (8.1)
ppe28	144	10 (6.9)
pe_pgrs38	30	2 (6.7)
ppe38	93	6 (6.5)
pe_pgrs16	79	5 (6.3)
ppe25	111	7 (6.3)
pe_pgrs36	16	1 (6.3)
ppe30	100	6 (6)
pe19	18	1 (5.6)
pe_pgrs18	39	2 (5.1)
pe_pgrs31	42	2 (4.8)
Ppe34	194	9 (4.6)
Ppe24	182	8 (4.4)

pe1	145	6 (4.1)
pe_pgrs63	106	4 (3.8)
ppe3	117	4 (3.4)
ppe18	92	3 (3.3)
pe_pgrs29	32	1 (3.1)
pe_pgrs45	32	1 (3.1)
pe_pgrs50	98	3 (3.1)
pe17	73	2 (2.7)
ppe53	75	2 (2.7)
pe_pgrs41	38	1 (2.6)
ppe68	79	2 (2.5)
pe8	46	1 (2.2)
ppe8	316	6 (1.9)
ppe43	107	2 (1.9)
ppe5	174	3 (1.7)
ppe1	117	2 (1.7)
ppe26	103	1 (1)
ppe11	105	1 (1)
ppe45	111	1 (0.9)
pe16	127	1 (0.8)
pe10	17	0 (0)
pe11	21	0 (0)
pe12	59	0 (0)
pe13	20	0 (0)
pe14	25	0 (0)
pe15	10	0 (0)
pe20	25	0 (0)
pe2	106	0 (0)
pe21	6	0 (0)
pe22	28	0 (0)
pe23	70	0 (0)
pe24	40	0 (0)
pe26	85	0 (0)
pe27	49	0 (0)
pe27A	2	0 (0)
pe28	25	0 (0)
pe29	10	0 (0)
pe31	13	0 (0)
pe32	16	0 (0)
pe33	13	0 (0)
pe34	22	0 (0)
pe35	9	0 (0)
pe36	7	0 (0)

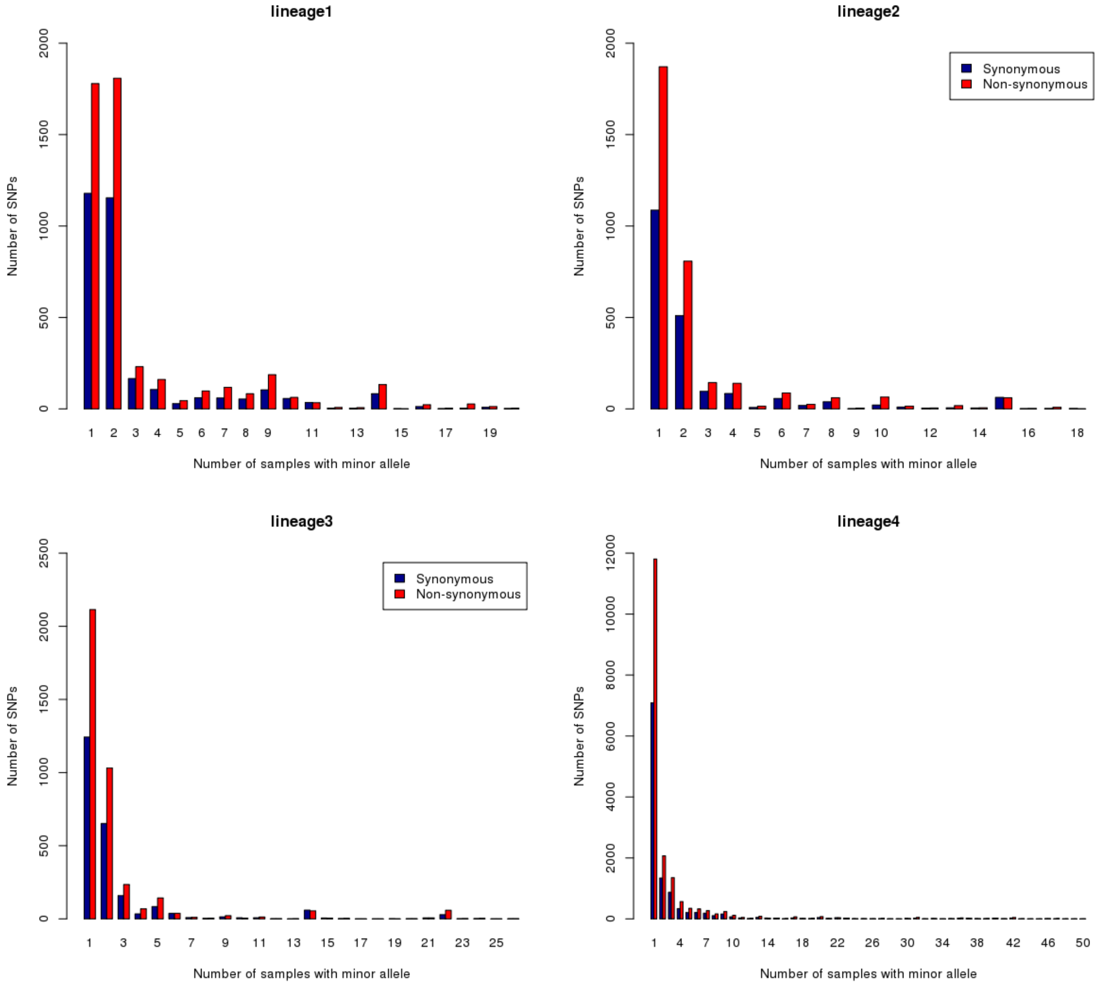
pe4	122	0 (0)
pe5	8	0 (0)
pe7	13	0 (0)
pe9	25	0 (0)
pe_pgrs11	113	0 (0)
pe_pgrs1	34	0 (0)
pe_pgrs14	48	0 (0)
pe_pgrs15	23	0 (0)
pe_pgrs17	31	0 (0)
pe_pgrs19	36	0 (0)
pe_pgrs21	35	0 (0)
pe_pgrs22	37	0 (0)
pe_pgrs2	31	0 (0)
pe_pgrs23	36	0 (0)
pe_pgrs24	36	0 (0)
pe_pgrs25	23	0 (0)
pe_pgrs27	26	0 (0)
pe_pgrs28	25	0 (0)
pe_pgrs30	111	0 (0)
pe_pgrs32	37	0 (0)
pe_pgrs33	33	0 (0)
pe_pgrs34	32	0 (0)
pe_pgrs35	98	0 (0)
pe_pgrs37	3	0 (0)
pe_pgrs39	48	0 (0)
pe_pgrs40	12	0 (0)
pe_pgrs42	30	0 (0)
pe_pgrs43	42	0 (0)
pe_pgrs4	38	0 (0)
pe_pgrs44	33	0 (0)
pe_pgrs46	30	0 (0)
pe_pgrs47	38	0 (0)
pe_pgrs48	17	0 (0)
pe_pgrs51	30	0 (0)
pe_pgrs52	25	0 (0)
pe_pgrs5	25	0 (0)
pe_pgrs53	27	0 (0)
pe_pgrs54	28	0 (0)
pe_pgrs55	32	0 (0)
pe_pgrs56	0	NA
pe_pgrs57	30	0 (0)
pe_pgrs58	28	0 (0)
pe_pgrs59	31	0 (0)

pe_pgrs61	5	0 (0)
pe_pgrs62	137	0 (0)
pe_pgrs6	41	0 (0)
pe_pgrs8	30	0 (0)
pe_pgrs9	41	0 (0)
ppe10	85	0 (0)
ppe12	71	0 (0)
ppe14	100	0 (0)
ppe15	115	0 (0)
ppe16	76	0 (0)
ppe17	93	0 (0)
ppe20	140	0 (0)
ppe2	131	0 (0)
ppe21	71	0 (0)
ppe23	96	0 (0)
ppe29	91	0 (0)
ppe31	98	0 (0)
ppe32	94	0 (0)
ppe33	83	0 (0)
ppe35	161	0 (0)
ppe36	49	0 (0)
ppe37	137	0 (0)
ppe39	22	0 (0)
ppe40	66	0 (0)
ppe41	49	0 (0)
ppe4	153	0 (0)
ppe42	116	0 (0)
ppe44	96	0 (0)
ppe48	29	0 (0)
ppe49	105	0 (0)
ppe50	45	0 (0)
ppe51	80	0 (0)
ppe55	374	0 (0)
ppe56	455	0 (0)
ppe58	46	0 (0)
ppe6	161	0 (0)
ppe61	93	0 (0)
ppe62	60	0 (0)
ppe63	134	0 (0)
ppe64	63	0 (0)
ppe66	79	0 (0)
ppe67	18	0 (0)
ppe69	63	0 (0)

ppe7	18	0 (0)
ppe9	45	0 (0)

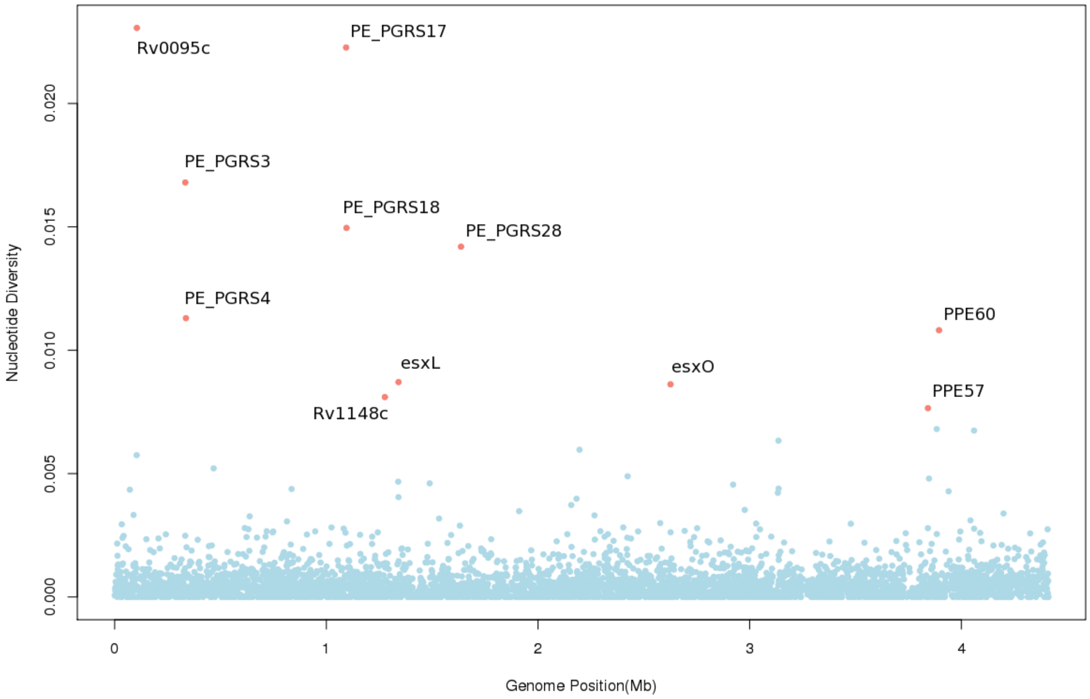
Supplementary File 2: Figure S1

Allele frequency spectra for each lineage by synonymous (blue) and non-synonymous (red) mutations. The peaks at intermediate allele frequencies include sub-lineage defining SNPs (Lineage 1 Indo-Oceanic; Lineage 2 East-Asian (Beijing); Lineage 3 East-African-Indian; Lineage 4 Euro-American).



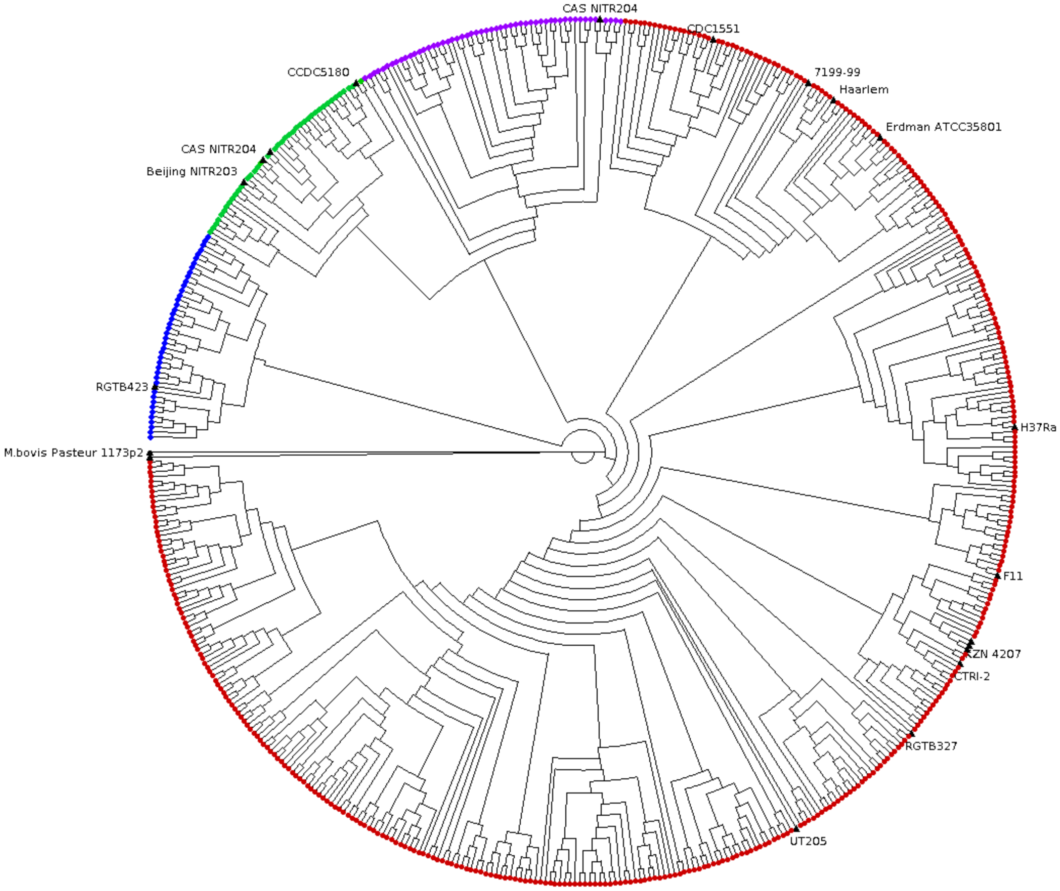
Supplementary file 3: Figure S2

Gene-based nucleotide diversity (π) for the 21 reference genomes. All genes with high nucleotide diversity ($\pi > 0.0075$) are labelled.



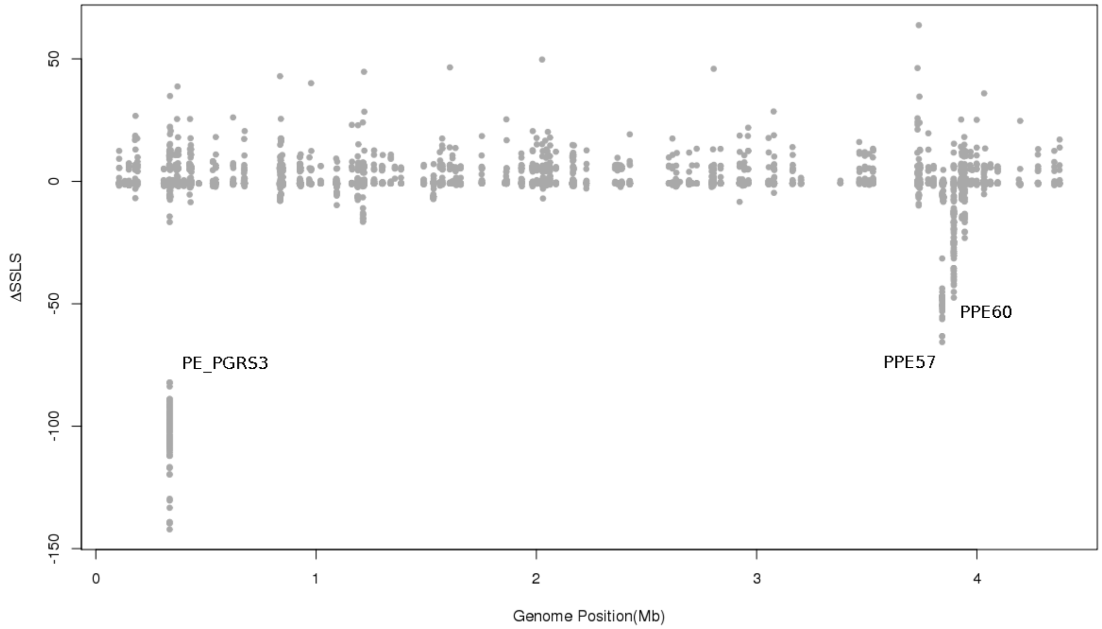
Supplementary file 4: Figure S3

Phylogenetic tree constructed using 50,540 genome-wide SNPs. Clear clustering according to lineage can be seen (Lineage 1 (Indo-Oceanic, green), lineage 2 (East-Asian (Beijing), blue), lineage 3 (East-African-Indian, purple), lineage 4 (Euro-American, red)). Reference genomes are labelled. M. canetti is annotated in cyan.



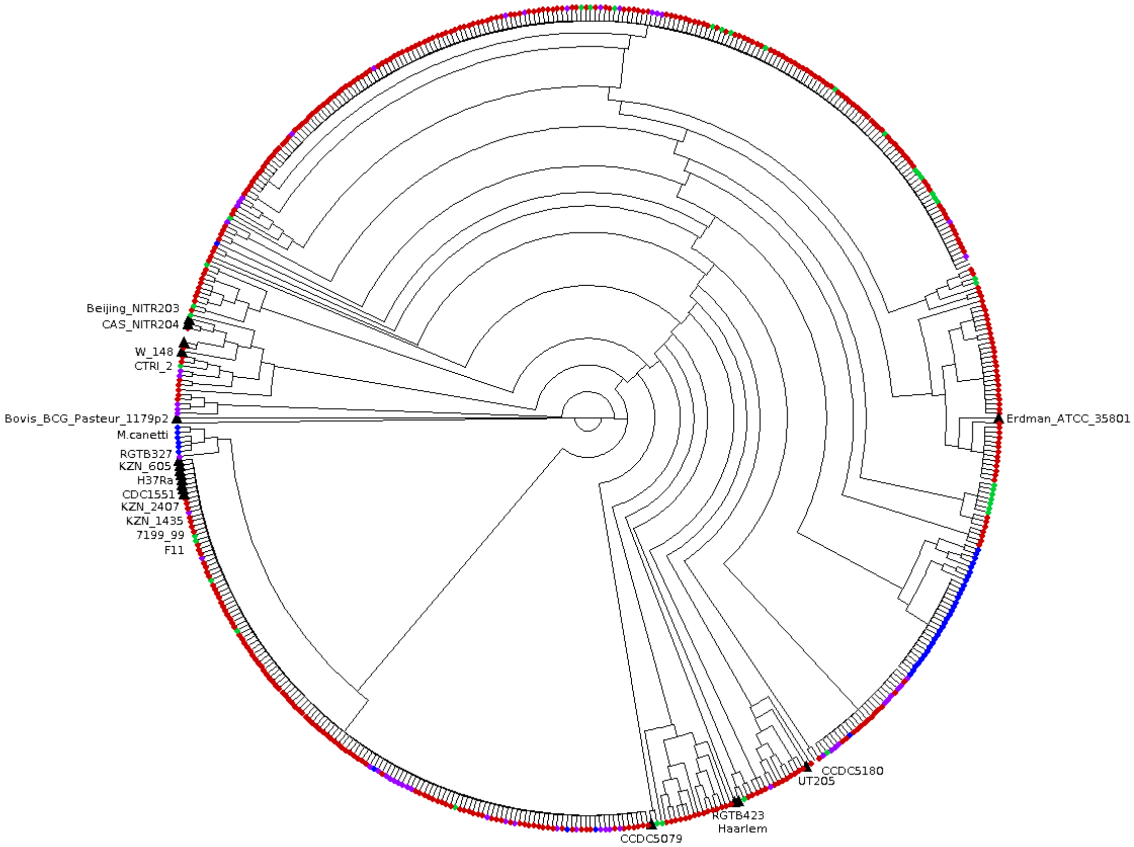
Supplementary file 5: Figure S4

Identifying sites leading to differences in tree topologies based on all SNPs (Additional file 4: Figure S3a) and only those from *pe/ppe* genes (Additional file 4: Figure S3b). The Δ Site wise log likelihood score (Δ SSLS) is calculated for each SNP in the *pe/ppe* gene alignments. Negative differences indicate SNP positions favouring the *pe/ppe* tree. SNPs in *pe_pgrs3, ppe57* and *ppe60* produce strong phylogenetic signals supporting the *pe/ppe* tree.



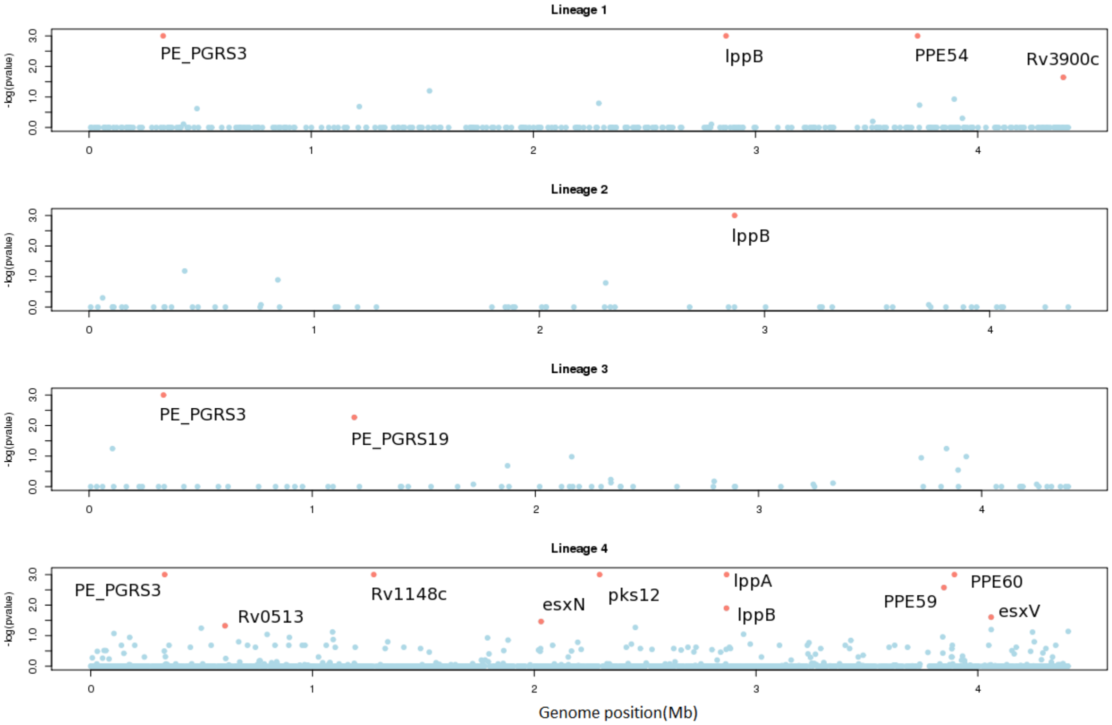
Supplementary file 6: Figure S5

Phylogenetic tree created using only SNPs from *pe_pgrs3*. No clear clustering by lineage is observed. However there are two major clades, one consistent with H37Rv (bottom-left).

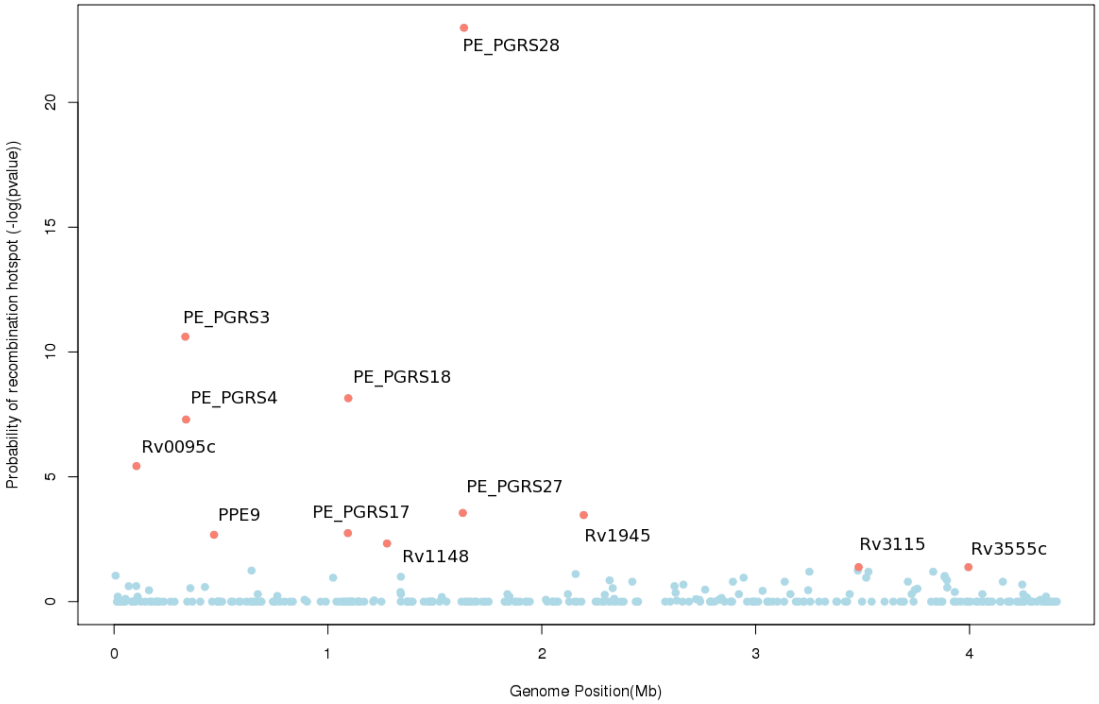


Supplementary file 7: Figure S6

Lineage-specific recombination hotspots. Manhattan plots showing genes that are likely to be recombination hotspots in each lineage (Lineage 1 Indo-Oceanic; Lineage 2 East-Asian (Beijing); Lineage 3 East-African-Indian; Lineage 4 Euro-American). The (−log10) p-value for the *phi* statistic is plotted against genome position. All genes with p-values < 0.05 are labelled.

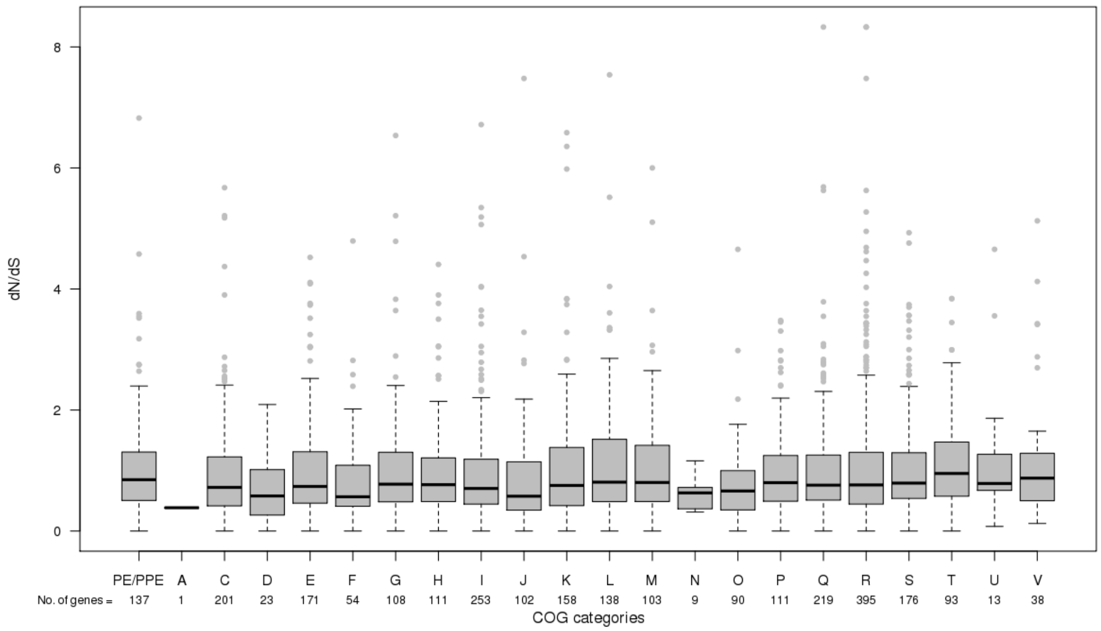


Supplementary file 8: Figure S7
Evidence of recombination at a gene level in the 21 reference genomes. A Manhattan plot showing genes that are likely to be recombination hotspots. The (−log10) p-value for the *phi* statistic is plotted against genome position. Genes with p-values less than 0.05 are shown.



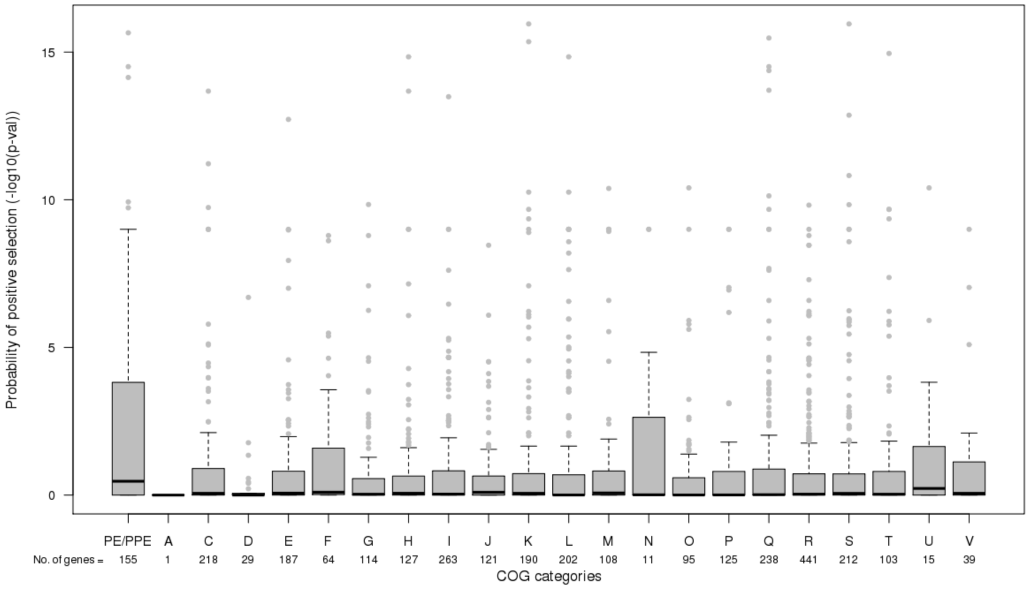
Supplementary file 9: Figure S8

Selection *dN/dS* values for each gene within Clusters of Orthologous Groups (COG*) categories. *ppe/N = *pe*/*ppe* genes annotated as COG category N, * COG categories: A RNA processing and modification, B Chromatin Structure and dynamics, C Energy production and conversion, D Cell cycle control and mitosis, E Amino Acid metabolism and transport, F Nucleotide metabolism and transport, G Carbohydrate metabolism and transport, H Coenzyme metabolism, ILipid metabolism, J Translation, K Transcription, L Replication and repair, M Cell wall/membrane/envelope biogenesis, N Cell motility, O Post-translational modification, protein turnover, chaperone functions, P Inorganic ion transport and metabolism, Q Secondary Structure, T Signal Transduction, U Intracellular trafficking and secretion, Y Nuclear structure, Z Cytoskeleton, RGeneral Functional Prediction only, S Function Unknown.



Supplementary file 10: Figure S9

Non-neutral evolution for genes within Clusters of Orthologous Groups (COG*) categories. Boxplots are constructed using (-log10) p-values of non-neutral evolution for each gene. *ppe/N = *pe/ppe* genes annotated as COG category N, * COG categories: A RNA processing and modification, B Chromatin Structure and dynamics, C Energy production and conversion, D Cell cycle control and mitosis, E Amino Acid metabolism and transport, F Nucleotide metabolism and transport, G Carbohydrate metabolism and transport, H Coenzyme metabolism, I Lipid metabolism, J Translation, KTranscription, L Replication and repair, M Cell wall/membrane/envelope biogenesis, N Cell motility, O Post-translational modification, protein turnover, chaperone functions, P Inorganic ion transport and metabolism, Q Secondary Structure, TSignal Transduction, U Intracellular trafficking and secretion, Y Nuclear structure, Z Cytoskeleton, R General Functional Prediction only, S Function Unknown.



RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Jody Phelan
Principal Supervisor	Taane Clark
Thesis Title	A Bioinformatic analysis of M. tuberculosis and host genomic data

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Scientific reports
Please list the paper's authors in the intended authorship order:	Jody Phelan, Paola Florez de Sessions, Leopold Tientcheu, Joao Perdigao, Diana Machado, Rumina Hasan, Zahra Hasan, Indra L. Bergval, Richard Anthony, Ruth McNerney, Martin Antonio, Isabel Portugal, Miguel Viveiros, Susana Campino, Martin L. Hibberd, Taane G Clark
Stage of publication	**In press**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I helped with the selection of the isolates to undergo PacBio sequencing. Upon receiving the raw sequence data, I set up the computing environment (SMRT portal) on our systems and used this to process the sequence data. I assembled the genomes and profiled the methylation using workflows available on the SMRT portal environment. After processing the data in SMRT portal, I wrote custom scripts to perform genome annotation, reference alignment, phylogenetic reconstruction, methylation profiles and pathway analyses. All custom analysis scripts were written in python and R by myself. I generated final plots in R. After finishing all analysis, I wrote the first draft of the manuscript and incorporated co-authors comments. I submitted to the journal and performed new analysis as required by

Student Signature: ____ **Date:** __________________

Supervisor Signature: ____ **Date:** __________________

Chapter 7

Methylation in *Mycobacterium tuberculosis* is lineage specific with associates mutations present globally

Methylation in *Mycobacterium tuberculosis* is lineage specific with associated mutations present globally

Jody Phelan[1], Paola Florez de Sessions[2,*], Leopold Tientcheu[1,3,*], Joao Perdigao [4,*], Diana Machado,[5*], Rumina Hasan[6], Zahra Hasan[6], Indra L. Bergval [7], Richard Anthony [7], Ruth McNerney [1,8], Martin Antonio [3], Isabel Portugal [4], Miguel Viveiros [5], Susana Campino[1], Martin L. Hibberd[1,2,**], Taane G Clark[1,9,**]

[1] Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom

[2] Genomics Institute Singapore, Singapore

[3] Vaccines and Immunity Theme, Medical Research Council Unit, The Gambia

[4] iMed.ULisboa - Research Institute for Medicines, Faculdade de Farmácia, Universidade de Lisboa, Portugal

[5] Unidade de Microbiologia Médica, Global Health and Tropical Medicine, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, UNL, Lisboa, Portugal

[6] Department of Pathology and Laboratory Medicine, The Aga Khan University, Stadium Road, P.O. Box 3500, Karachi 74800, Pakistan

[7] Royal Tropical Institute, KIT Biomedical Research, Meibergdreef 39, 1105 AZ Amsterdam, The Netherlands

[8] Lung Infection and Immunity Unit, UCT Lung Institute, University of Cape Town, Groote Schuur Hospital, Observatory, 7925, Cape Town, South Africa.

[9] Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom

* joint authors

** joint corresponding authors

Abstract

DNA methylation is an epigenetic modification of the genome involved in regulating crucial cellular processes, including transcription and chromosome stability. Advances in PacBio sequencing technologies can be used to robustly reveal methylation sites. The methylome of the *Mycobacterium tuberculosis* complex is poorly understood but may be involved in virulence, hypoxic survival and the emergence of drug resistance. In the most extensive study to date, we characterise the methylome across the 4 major lineages of *M. tuberculosis* and 2 lineages of *M. africanum,* the leading causes of tuberculosis disease in humans. We reveal lineage-specific methylated motifs and strain-specific mutations that are abundant globally and likely to explain loss of function in the respective methyltransferases. Our work provides a set of sixteen new complete reference genomes for the *Mycobacterium tuberculosis* complex, including complete lineage 5 genomes. Insights into lineage-specific methylomes will further elucidate underlying biological mechanisms and other important phenotypes of the epi-genome.

Introduction

Tuberculosis disease (TB) caused by pathogens of the *Mycobacterium tuberculosis* complex are an important global public health issue worldwide, with >9 million new cases and 1.7 million deaths each year[1]. A combination of the increasing prevalence of anti-tuberculosis drug resistance, HIV/AIDS infection interaction, and an under-equipped arsenal - requiring new effective treatments and vaccines, are a major barrier to disease control. The *M. tuberculosis* genome (size 4.4Mb, GC-content 60%) is characterised by low sequence diversity[2], with known variation between stain-types, including between three 'ancient' (1, 5, 6), three 'modern' (2, 3, 4), and one intermediate lineage (7)[3]. The lineages vary in propensity to transmit and cause disease[4]; with modern strain lineages, including Beijing strains, being more successful in terms of their geographical spread and have a shorter latency in humans[5]. However, results are inconsistent and there is considerable inter-strain variation within lineages, which is difficult to explain in the context of the low sequence diversity[6].

Several lines of evidence have revealed N6-methyladenine (^{m6}A) and 5-methylcytosine (^{m5}C) methylation mechanisms within *M. tuberculosis* genomes. Motifs within three DNA methyltransferases (MTases), *mamA, mamB, hdsS.1, hsdM*, and *hsdS* are responsible for m6A modification[7,8]. *MamA* also influences gene expression in *M. tuberculosis* and plays an important but strain-specific role in fitness during hypoxia, promoting survival in discrete host microenvironments[7]. Genetic and potentially transcriptomic differences, may play important roles in determining the clinical outcome differences observed between these strains. Genetic differences may be further modified by epigenetic mechanisms, as observed in other bacterial species[9], however methylome data has been rarely considered for the *M. tuberculosis* complex. Here we present to our knowledge the largest and most diverse study of methylation in *M. tuberculosis* using PacBio technology, and identify key mutations in associated genes, which appear to be present across a phylogeny based on a global set of isolates.

RESULTS

New reference genomes

Sixteen samples representing the lineages 1, 2, 4, 5 and 6 were sequenced on the PacBio platform (**Supplementary table 1,** n=16), and supplemented by raw sequence data for a lineage 3 strain and a H37Rv strain (CHIN_F1) from earlier work (n=2)[8]. High quality assemblies (no. contigs <10) were generated for the 18 isolates, with most isolates assembled into one contig (median n50 = 4.38Mb, median genome length = 4.42Mb). After aligning to the H37Rv reference, we found 10,353 unique small variant sites, with 50.7% of positions having alternate alleles in only one sample. A maximum likelihood tree was constructed using the variants (**Figure 1**) and demonstrated the expected clustering by lineage, with two lineage 1 strains (WBB1008_SL1975, WBB1007_LQ1975) being near identical.

The error rate in the PacBio consensus sequences was assessed in three isolates (**Supplementary table 1;** WBB446_ARS7884 (LAM strain, lineage 4), WBB448_HPV115_08 (LAM strain, lineage 4), WBB445_ARS7496 (Beijing strain, lineage 2)) that also had Illumina short read data with high coverage genome-wide (>50-fold). Alignment of the short reads to the consensus sequences revealed low numbers of discordant SNPs (range: 0-6), but slightly higher numbers of discordant insertions and deletions (indels) (range: 2 to 26) due to incorrect assembly at homopolymeric sites in the genome. More generally, further analysis of these isolates revealed the advantages of using lineage-specific reference genomes. First, using sets of 100 independent strains in each lineage[2], there was a marginal improvement in the number of reads mapped compared to using an alignment to H37Rv (mean increase: lineage 1 0.47%, lineage 2 0.33%, lineage 3 0.25%). As *M. tuberculosis* has very clonal genome, most of the genome shares near 100% identity across lineages, and therefore large improvements in overall mappability would not be expected. Second, we considered strain-specific regions in the highly variable *PE_PGRS3/4* and *PE_PGRS17/18* genes, which were hypothesised from *de novo* assembly analysis to have undergone a large genomic rearrangement in Beijing strains[10]. The PacBio consensus sequence confirmed the large re-arrangements in WBB445_ARS7496 (Beijing). These re-arrangements could be identified in coverage profiles through mapping WBB445_ARS7496 short reads to its own PacBio consensus sequence, but not to the H37Rv reference or other non-Beijing study consensus sequences (**Supplementary Figure 1**).

Annotation of the new reference genomes using *prokka* software[11], guided by H37Rv protein sequences, revealed differences in the number of genes (range: 4028 to 4217). The CHIN_F1 strain (H37Rv) had a greater number of inferred genes (4217) than the H37Rv reference (ASM19595v2, 4093 genes), which may indicate that the automatic annotation software could be over-estimating numbers of genes. However, overall, there was a high degree of conservation among isolates across orthologous groups of genes (3666/4250, 86%). Hierarchical clustering of isolates using the number of shared orthogroups as a metric of genetic distance, revealed expected lineage-specific clustering, except for the CHIN_F1 strain which clustered outside lineage 4 and closer to lineage 3 (**Supplementary figure 2**).

Methylation motif analysis

Using the Modification and Motif Analysis pipeline in the SMRT portal (https://github.com/PacificBiosciences/SMRT-Analysis), Pacbio sequence data can be used to robustly reveal methylation sites. A variable number of motifs (range: 3-13) were found per isolate, with 45 unique motifs discovered across the entire dataset of 18 isolates. Three high quality methylated motifs (quality value score >100) were detected across almost all isolates: CACGCAG (17/18 isolates), GATN4RTAC (14/18), and CTCCAG (15/18) (**Table 1**). Partner motifs for GATN4RTAC and CTCCAG were also found indicating methylation on both the forward and reverse strand, while CACGCAG is only hemi-methylated as no partner motif was found. These motifs have previously been reported[8,12]. The number of occurrences of each motif was found to vary slightly across isolates (range: GATN4RTAC 349-366, CACGCAG 811-828, CTCCAG 1928-1957).

By considering the motifs across all isolate genome assemblies and inspection of the raw inter-pulse duration (IPD) ratios at each nucleotide position in the motif, we found that isolates where the motif was present but had no evidence of modification across nucleotides (**Supplementary figure 3**). There was some variability across and within strain types in the percent of motifs methylated. In particular, although motifs were mostly close to 100% (or alternatively 0%) methylated, three isolates had a substantially different percentage for the CACGCAG motif (median (range) %: 60.0 (52.5-63.7)) (**Table 2**). Methylation of the other two motifs (GATN4RTAC, CACGCAG) did not seem affected in these isolates (range 93.9 - 99.3%).

To explain the differences in methylation pattern we identified mutations in methyltransferase genes that have been associated with each motif (GATN4RTAC: *hsdS.1*, *hsdM* and *hsdS;* CTCCAG: *mamA*; CACGCAG: *mamB*)[8] (**Table 2**). In particular, we scanned for mutations that were present in methylation-deficient isolates, as identified through analysis of PacBio data, which could putatively explain loss of function in the respective methyltransferase. For the GATN4RTAC motif we found three unique mutations in four isolates with an absence of methylation, confirming those identified in previous reports[8]. Three methylation-absent isolates had the presence of the *hsdM* P306L mutation. Additionally, one sample had two mutations which were not present in any other isolates: *hsdM* G173D and *hsdS* L119R. Three samples did not exhibit any methylation at the CTCCAG motif, and we identified three unique mutations in *mamA*, one of which was present in two samples. One isolate had an E270A mutation and frameshift deletion at position 1257, however through phylogenetic ancestral reconstruction we deduced that the E270A mutation occurred before the deletion (**Figure 1**). The two other isolates had E270A and previously uncharacterised A460T mutations, respectively. For the CACGCAG motif, the CHIN_F1 strain has a truncated *mamB* gene which has been reported elsewhere[8], and verified here. Additionally, we found all three lineage 1 strains, which exhibited ~50-60% methylation, to have a novel S253L mutation in *mamB*.

Pathway analysis

To look for the non-random association of methylation sites and protein families or biological pathways we performed a pathway analysis using DAVID software[13]. Each of the three motifs was considered individually. Motifs were associated with genes based on overlap with an annotated coding region or the closest promoter. Most motifs were found in the coding regions, with few found within promoters (defined as the 50 nucleotides before a start codon) (**Supplementary Figure 4**). For GATN4RTAC, we found an enrichment of cell membrane associated genes (Bonferroni corrected P-value (P*) = 0.021) and plasma associated genes (P* = 0.023) in motif-containing genes compared to genes without the motifs. For CTCCAG, motif-containing genes were enriched for nucleotide binding (P* = 9.99e-13) and cell wall (P* = 1.63e-5) among others (**Supplementary table 2**). For the CACGCAG motif we found several enriched pathways involved in fatty acid and polykeytide synthesis (P* = 9.26E-05) among others. DAVID software was used to test whether there was targeted absence of methylation of genes in a specific pathway. Genes with an absence of methylation in excess of 60% of the isolates were compared against all *M. tuberculosis* genes to look for enrichment of specific pathways. This analysis was performed on an overall and per-lineage basis. No pathways reported significant results (P*>0.05). When comparing motif-containing unmethylated to motif-containing methylated genes on a lineage basis we did not find any significantly enriched pathways, although the small number of isolates is likely to lead to reduced power to detect true enriched pathways.

Motifs in a global context

To describe the six mutations we identified as affecting methylation in a global context, we analysed a large collection (n = 6465) of isolates representing lineages 1 (9.5%), 2 (15.8%), 3 (15.4%) and 4 (59.3%). We also analysed lineage 5 (n=4) strains and lineage 6 (n=26) strains, a combination of our own data and those described elsewhere[14]. We found five of the six mutations identified above in the global dataset, occurring predominantly in single lineages with low frequencies in other lineages (**Table 2**), and originating at unique positions in the phylogeny (**Figure 2**). None of the six mutations were found in the lineage 5/6 dataset, except for the isolate in which we originally found the *mamA* 460T mutation. The *mamA* A460T is likely to be specific to a subclade of lineage 6. Three mutations affecting the GATN4RTAC motif were found at high allele frequency (*hsdM* G173D: 0.15, *hsdM* P306L: 0.42, *hsdS* L119R: 0.15) and affected ~57% of the isolates. The *hsdM* P306L mutation is a phylogenetically deep mutation which occurs in a sub-clade of lineages 4.3 to 4.9 (H3, H4, LAM, LAM1, LAM10-CAM, LAM11-ZWE, LAM3, LAM4, LAM9, S, T1, T2, T2-Uganda, T3, T4, T5). The *hsdM* G173D and *hsdS* L119R mutations are present in all lineage 3 isolates. The *mamB* S253L mutation affecting the CACGCAG motif is present only in a subclade of lineage 1 (EAI6). The *mamA* E270A mutation affecting the CTCCAG motif is present in all lineage 2 strains. Assuming that these mutations do indeed cause the absence of methylation on the genome there is a stark difference between the motifs in the lineages and number of samples which have active methylation.

DISCUSSION

We have presented 16 new reference genomes and methylomes of strains with diverse genetic backgrounds. The ability of PacBio technology to produce long reads leads to complete genome assemblies that capture both small and large genomic variations and have a very high accuracy at repetitive regions such as the *pe/ppe* genes. Most whole genome sequencing projects have focused on lineages 1 to 4 because of their prevalence and global distribution, however recent studies have shown a large amount of genetic diversity to be present within lineages 5 and 6[14]. Additionally, an intriguing question remains why lineages 5 and 6 are localised to West Africa and have not spread globally. The lineage specific variants and differences in gene content (including the pe*/ppe* genes) reported here, building on previous work[3], could potentially play a role in specific host population adaptation. We present, to our knowledge, the first complete lineage 5 reference genomes, and increase substantially the number of lineage 6 reference genomes available. These references will be useful in future whole genome sequencing projects that investigate the genetic diversity of lineage 5 and 6 strains, as well as strain-host genetic interactions. By aligning Illumina reads to our references we find there to be a small increase in the number of reads mapping (0.25-0.47%), particularly in genomic regions where sequences are either not present or highly variable in the H37Rv reference. By performing automatic annotation and clustering of protein sequence into clusters of orthologues, we report a significant difference in the gene content between strains. Overall, these new reference sequences could serve to improve the accuracy of resequencing experiments by facilitating lineage-specific mapping at highly variable regions and to improve our understanding of large structural variations such as novel insertions, as well as rearrangements between lineages.

The PacBio technology allowed us to characterise the methylation at sites along the genome. Across the 18 isolates, three motifs are methylated to varying degrees. While most isolates had close to 100% methylation with an active MTase, the three lineage 1 isolates had 53-64% methylation at the CACGCAG motif while maintaining near 100% methylation on both other motifs. We identified a number of mutations which associate with the absence of methylation, some of which have been reported before[7,8]. Five of these mutations were present in a large global phylogeny consisting of *M. tuberculosis* lineages (1-4) strains. The frequency of the potential loss of function mutations is reasonably high. For example, the three mutations (*hsdM* G173D, *hsdM* P306L and *hsdS* L119R) affecting the GATN4RTAC motif methylation were present in all available lineage 3 (all sub-lineages) strains, as well as across a larger number of lineage 4 sub-lineages (including H3, H4, LAM, LAM1, LAM10-CAM, LAM11-ZWE, LAM3, LAM4, LAM9, S, T1, T2, T2-Uganda, T3, T4, T5), but absent in other lineages. Similarly, the other motifs (CTCCAG and CACGCAG) have a lower frequency of loss of function mutations, but are also strain specific. Follow-up investigation is required to provide an insight into the essential and functional nature of methylation, and its association with the different motifs. Interestingly the lineage 2 strains, which have been reported to be highly virulent[15], lack methylation in the most abundant motif (CTCCAG) putatively due to the *mamA* E270A mutation. Differential methylation patterns could provide a possible explanation for the increased virulence in this clade, as genetic distance is relatively small. Similarly, the *mamB* S253L mutation related to the CACGCAG motif seems only present in EAI6 strains, and whilst little is known whether these strains are more virulent than other lineage 1 "ancient" strains, they have spread globally and have been associated with recent outbreaks[16], unlike other lineage 1 strains.

It has been hypothesised that DNA methylation influences transcription[9] and therefore it would be expected to see a differences in transcriptional profiles of genes where there is differential methylation. Additionally, although no correlation was found with drug resistance (data not shown), transcriptional regulation by DNA methylation could potentially contribute towards observed strain-specific differences in the acquisition of mutations involved in drug resistance[17]. Whilst, our work has shed new light on *M. tuberculosis* methylation, future work should consider more diverse strains and integrate transcriptomic data to further elucidate underlying biological mechanisms and associating them with virulence and other important phenotypic outcomes including antibiotic resistance.

MATERIALS AND METHODS

Samples and SMRT sequencing

DNA was extracted from *M. tuberculosis* cultures of clinical samples, processed using methods described elsewhere[2,3,18]. Samples were sequenced using Pacific Biosciences (PacBio) RSII long read technology. Additionally, raw data for two isolates was downloaded from the SRA project SRP064893 to be included in the current study. All raw sequencing data are available, and the study accession numbers are listed in **Supplementary table 1.**

Bioinformatic analysis

Sequencing reads were assembled using Hierarchical Genome Assembly Process HGAP2 implemented in the SMRT Portal software suite. Short low confidence contigs (length<1000 or identity < 90%) were removed from subsequent analyses. Overlap between the start and end of large contigs were found by self-aligning using *Mummer* software (mummer.sourceforge.net) and removed using in-house scripts. Contigs were aligned, scaffolds inferred, reordered and, if needed, reverse-complemented according to the H37Rv reference using the *mummer* tool and in-house scripts. Following this the reads were realigned to the scaffolds to improve the consensus concordance. The final consensus genome for each sample was annotated using *prokka* automatic annotation tool[11] using the H37Rv protein sequences to annotate the genes found. *Mummer* software was used to align the consensus against H37Rv to identify small variants (SNPs and indels). Methylation analysis was performed using the Modification and Motif Analysis pipeline in SMRT portal, and outputted motifs of interest. All high-quality motifs were used in further downstream analysis. A maximum likelihood phylogenetic tree was built using RAxML with all polymorphic SNP sites found. Pathway analysis was performed by assigning a gene to each motif found in a genome. Genes were assigned using overlap with the coding region or promoter of a gene. Statistical enrichment analysis was performed using *DAVID* software[14] and compared: (i) all motif-containing genes to all *M. tuberculosis* genes; (ii) all un-methylated genes to all motif-containing genes. To identify mutations within lineages 5 and 6, genome assemblies were downloaded from *genbank*[14] and aligned to the H37Rv reference using the *mummer* tool with default parameters. Variants were then called using the *snp-snps* algorithm, with the "-C" parameter invoked, leading to the reporting of variants from unambiguous alignments.

REFERENCES

1. Organisation, W. H. Global tuberculosis report 2016. (2016).
2. Coll, F. *et al.* PolyTB: A genomic variation map for Mycobacterium tuberculosis. *Tuberculosis* **94,** 346–354 (2014).
3. Coll, F. *et al.* A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nat. Commun.* **5,** 4812 (2014).
4. Guerra-Assunção, J. *et al.* Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *Elife* **4,** (2015).
5. Gagneux, S. Host-pathogen coevolution in human tuberculosis. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367,** 850–9 (2012).
6. Portevin, D., Gagneux, S., Comas, I., Young, D. & Belardelli, F. Human Macrophage Responses to Clinical Isolates from the Mycobacterium tuberculosis Complex Discriminate between Ancient and Modern Lineages. *PLoS Pathog.* **7,** e1001307 (2011).
7. Shell, S. S. *et al.* DNA Methylation Impacts Gene Expression and Ensures Hypoxic Survival of Mycobacterium tuberculosis. *PLoS Pathog.* **9,** e1003419 (2013).

8. Zhu, L. *et al.* Precision methylome characterization of *Mycobacterium tuberculosis* complex (MTBC) using PacBio single-molecule real-time (SMRT) technology. *Nucleic Acids Res.* **44,** 730–743 (2016).

9. Casadesús, J. & Low, D. Epigenetic gene regulation in the bacterial world. *Microbiol. Mol. Biol. Rev.* **70,** 830–56 (2006).

10. Phelan, J. E. *et al.* Recombination in pe/ppe genes contributes to genetic variation in Mycobacterium tuberculosis lineages. *BMC Genomics* **17,** 151 (2016).

11. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068–2069 (2014).

12. Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. REBASE: restriction enzymes and methyltransferases. *Nucleic Acids Res.* **31,** 418–20 (2003).

13. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57 (2008).

14. Winglee, K. *et al.* Whole Genome Sequencing of Mycobacterium africanum Strains from Mali Provides Insights into the Mechanisms of Geographic Restriction. *PLoS Negl. Trop. Dis.* **10,** e0004332 (2016).

15. Reiling, N. *et al.* Clade-specific virulence patterns of Mycobacterium tuberculosis complex strains in human primary macrophages and aerogenically infected mice. *MBio* **4,** e00250-13 (2013).

16. Duarte, T. A. *et al.* A systematic review of East African-Indian family of Mycobacterium tuberculosis in Brazil. *Brazilian J. Infect. Dis.* **21,** 317–324 (2017).

17. Ford, C. B. *et al.* Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat. Genet.* **45,** 784–90 (2013).

18. Coll, F. *et al.* The Mycobacterium tuberculosis resistome from a genome-wide analysis of multi- and extensively drug-resistant tuberculosis. *Under Rev.* (2017).

ACKNOWLEDGMENTS

JPh is funded by a BBSRC PhD studentship. TGC is funded by the Medical Research Council UK (Grant no. MR/K000551/1 and MR/M01360X/1, MR/N010469/1, MC_PC_15103). SC is funded by Medical Research Council UK grants (MR/M01360X/1, MC_PC_15103). We gratefully acknowledge the Scientific Computing Group for data management and compute infrastructure at Genome Institute of Singapore for their help. The MRC eMedLab computing resource was used for bioinformatics and statistical analysis. The authors declare no conflicts of interest.

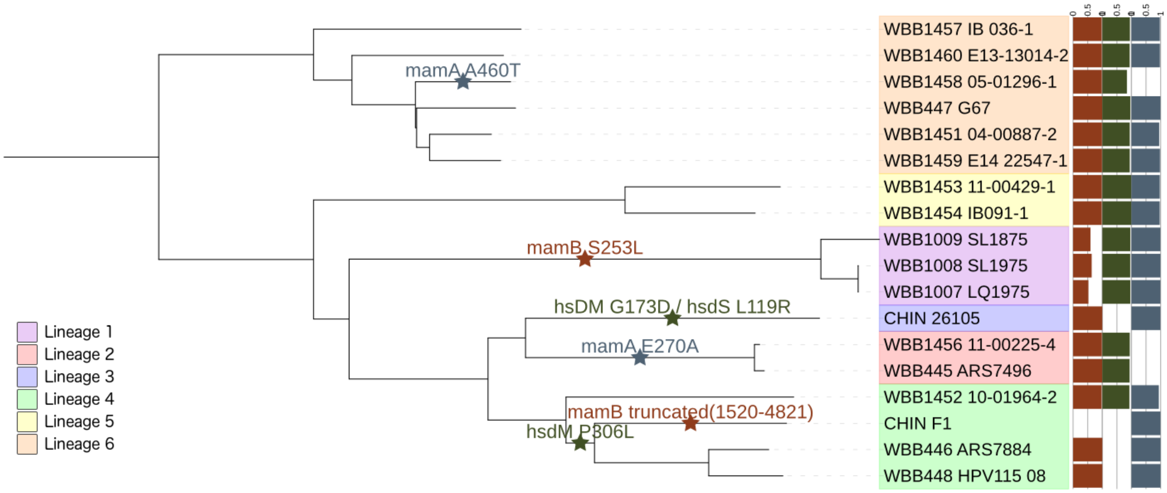
AUTHOR CONTRIBUTIONS

MLH and TGC conceived and directed the project. RM, SC and TGC coordinated sample collection. LT, JP, DM, RH, ZH, ILB, RA, IP and MV undertook sample collection and DNA extraction. PFdS and MLH coordinated sequencing. JPh performed bioinformatic and statistical analyses under the supervision of MLH and TGC. JPh, PFdS, SC, MLH and TGC interpreted results. JPh, MLH and TGC wrote the first draft of the manuscript. All authors commented and edited on various versions of the draft manuscript and approved the final manuscript. JPh, MLH and TGC compiled the final manuscript.

DISCLOSURE DECLARATION

The authors declare that they have no conflicts of interest.

Figure 1
Phylogeny of the *Mycobacterium tuberculosis* complex isolate consensus sequences (n=18) annotated with loss of function mutations in MTase genes



Motif	Mutation	Lineage 1 (n=617)	Lineage 2 (n=1021)	Lineage 3 (n=993)	Lineage 4 (n=3834)	Lineage 5 (n=4)	Lineage 6 (n=26)
GATN4RTAC	*hsdM* P306L	-	-	-	**0.71****	-	-
GATN4RTAC	*hsdM* G173D	-	-	**1**	-	-	-
GATN4RTAC	*hsdS* L119R	-	-	**1**	-	-	-
CACGCAG	*mamB* S253L	**0.23***	-	-	-	-	-
CTCCAG	*mamA* E270A	-	**1**	-	-	-	-
CTCCAG	*mamA* A460T	-	-	-	-	-	**0.04**

A maximum likelihood phylogenetic tree, with the % of methylated motifs and potential loss of function mutations in MTase genes annotated. Allele frequencies of putative methylation related mutations across a global collection of *M. tuberculosis* isolates; * EAI6 stains, ** lineages 4.3 to 4.9, - indicates absence

Figure 2
Five methylation-affecting mutations in a global collection of isolates (n = 6465; lineage 1 617 (9.5%), lineage 2 1021 (15.8%), lineage 3 993 (15.4%), lineage 4 3834 (59.3%)[18])

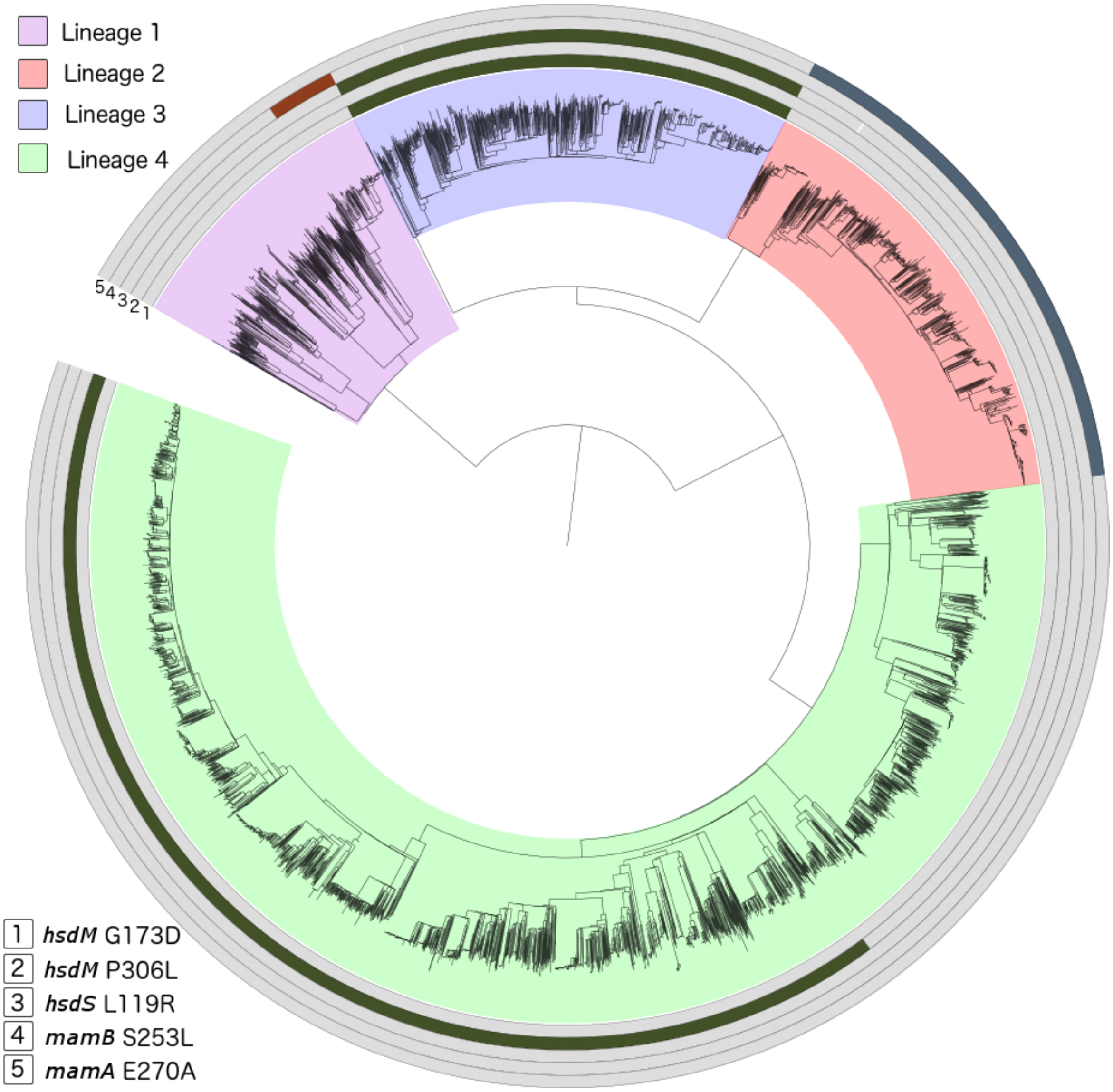
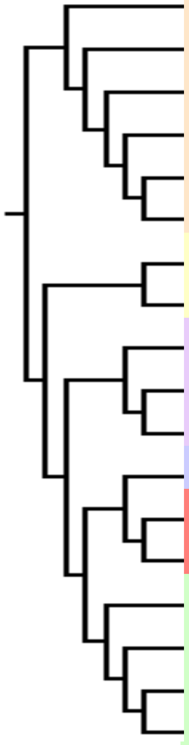


Table 1
Methylation of motifs and their proportion in the genome sequence assemblies of each isolate.



Isolate [lineage]	CACGCAG	GATNNNNRTAC	GTAYNNNNATC	CTCCAG	CTGGAG
WBB1457_IB_036-1 [6]	793/811 (0.98)	332/351 (0.95)	332/351 (0.95)	1885/1934 (0.97)	1828/1934 (0.95)
WBB1460_E13-13014-2 [6]	799/813 (0.98)	328/350 (0.94)	327/350 (0.93)	1892/1937 (0.98)	1825/1937 (0.94)
WBB1458_05-01296-1 [6]	799/813 (0.98)	294/349 (0.84)	290/349 (0.83)	0/1932 (0.00)	0/1932 (0.00)
WBB447_G67 [6]	814/814 (1.00)	338/352 (0.96)	336/352 (0.95)	1923/1933 (0.99)	1922/1933 (0.99)
WBB1451_04-00887-2 [6]	802/812 (0.99)	328/349 (0.94)	325/349 (0.93)	1842/1933 (0.95)	1801/1933 (0.93)
WBB1459_E14_22547-1 [6]	802/811 (0.99)	328/349 (0.94)	329/349 (0.94)	1891/1934 (0.98)	1833/1934 (0.95)
WBB1453_11-00429-1 [5]	814/828 (0.98)	357/362 (0.99)	355/362 (0.98)	1889/1942 (0.97)	1825/1942 (0.94)
WBB1454_IB091-1 [5]	807/823 (0.98)	356/358 (0.99)	353/358 (0.99)	1874/1929 (0.97)	1819/1929 (0.94)
WBB1009_SL1875 [1]	492/826 (0.60)	345/360 (0.96)	341/360 (0.95)	1942/1957 (0.99)	1885/1957 (0.96)
WBB1008_SL1975 [1]	526/826 (0.64)	344/360 (0.96)	345/360 (0.96)	1945/1956 (0.99)	1906/1956 (0.97)
WBB1007_LQ1975 [1]	434/826 (0.53)	345/360 (0.96)	338/360 (0.94)	1943/1956 (0.99)	1893/1956 (0.97)
CHIN_26105 [3]	823/824 (1.00)	0/362 (0.00)	0/362 (0.00)	1939/1954 (0.99)	1942/1954 (0.99)
WBB1456_11-00225-4 [2]	813/826 (0.98)	344/366 (0.94)	349/366 (0.95)	0/1949 (0.00)	0/1949 (0.00)
WBB445_ARS7496 [2]	824/824 (1.00)	339/363 (0.93)	340/363 (0.94)	0/1947 (0.00)	0/1947 (0.00)
WBB1452_10-01964-2 [4]	798/817 (0.98)	332/358 (0.93)	321/358 (0.90)	1828/1947 (0.94)	1748/1947 (0.90)
CHIN_F1 [4]	0/820 (0.00)	0/361 (0.00)	0/361 (0.00)	1937/1948 (0.99)	1937/1948 (0.99)
WBB446_ARS7884 [4]	817/817 (1.00)	0/357 (0.00)	0/357 (0.00)	1932/1933 (1.00)	1927/1933 (1.00)
WBB448_HPV115_08 [4]	814/814 (1.00)	0/355 (0.00)	0/355 (0.00)	1927/1928 (1.00)	1924/1928 (1.00)

The phylogenetic relationship and fraction of motifs methylated for each strain. Most values are close to either 0.95 or 0 indicating the presence or complete absence of methylation, however, all lineage 1 strains had approximately half of their CACGCAG motif methylated

Supplementary table 1
The isolates analysed

Isolate ID	Country	Source	N50	Num. Contigs	Genome length	Lineage	Sub-lineage	SRA accession
WBB1007_LQ1975	Mozambique	Sequenced	4450176	1	4450176	1	1.1.3 (EAI6)	PRJEB21888
WBB1008_SL1975	Mozambique	Sequenced	4467776	1	4467776	1	1.1.3 (EAI6)	PRJEB21888
WBB1009_SL1875	Mozambique	Sequenced	4438486	1	4438486	1	1.1.3 (EAI6)	PRJEB21888
WBB1456_11-00225-4	Gambia	Sequenced	4415343	1	4415343	2	2.2.1 (Beijing)*	PRJEB21888
WBB445_ARS7496	**Portugal**	**Sequenced**	**4415871**	**3**	**4446789**	**2**	**2.2.1 (Beijing)***	PRJEB21888
CHIN_26105	China	SRA	4440106	1	4440106	3	3 (CAS)*	SRP064893
WBB1452_10-01964-2	Gambia	Sequenced	4416076	2	4430073	4	4.1.2.1 (Haarlem)*	PRJEB21888
WBB446_ARS7884	**Portugal**	**Sequenced**	**4375931**	**3**	**4396369**	**4**	**4.3.4.2 (LAM)***	PRJEB21888
WBB448_HPV115_08	**Portugal**	**Sequenced**	**4385381**	**1**	**4385381**	**4**	**4.3.4.2 (LAM)***	PRJEB21888
CHIN_F1	China	SRA	4125500	5	4438875	4	4.9 (T1-H37Rv)	SRP064893
WBB1453_11-00429-1	Gambia	Sequenced	4430643	1	4430643	5	5 (Afr2/3)	PRJEB21888
WBB1454_IB091-1	Nigeria	Sequenced	3865667	3	4419358	5	5 (Afr2/3)	PRJEB21888
WBB1451_04-00887-2	Gambia	Sequenced	716074	6	4393399	6	6 (Afr1)	PRJEB21888
WBB1457_IB_036-1	Nigeria	Sequenced	2521417	4	4387174	6	6 (Afr1)	PRJEB21888
WBB1458_05-01296-1	Gambia	Sequenced	2446180	2	4369685	6	6 (Afr1)	PRJEB21888
WBB1459_E14_22547-1	Gambia	Sequenced	4382305	2	4384418	6	6 (Afr1)	PRJEB21888
WBB1460_E13-13014-2	Gambia	Sequenced	2963146	4	4413823	6	6 (Afr1)	PRJEB21888
WBB447_G67	Guinea-Bissau	Sequenced	2330737	3	4388314	6	6 (Afr1)	PRJEB21888

Bolded isolates also have Illumina short read data; Sub-lineages inferred using barcoding SNPs[3]; Afr = *M. africanum;* * known to be highly virulent[5]; SRA short read archive

Supplementary table 2
Pathway analysis of genes containing motifs

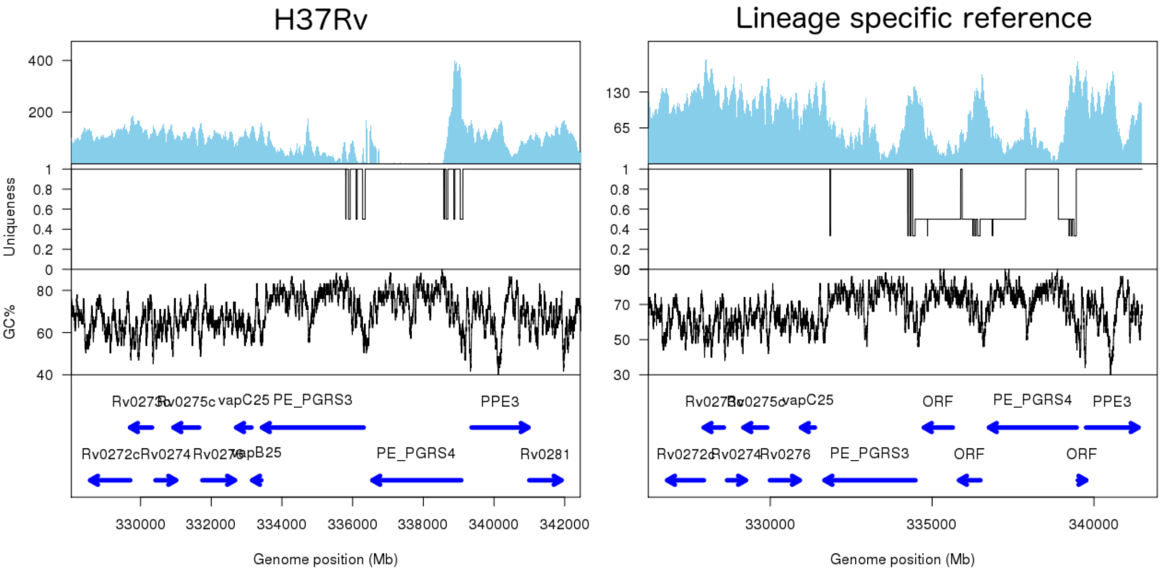
Motif	Ontological annotation	Count	% of genes in pathway	Fold Enrichment	P-value*
CTCCAG	ATP-binding	178	12.4	1.6	2.85E-14
CTCCAG	Cell wall	327	22.8	1.2	1.90E-05
CTCCAG	Plasma membrane	632	44	1.1	1.13E-04
CTCCAG	Phosphoprotein	52	3.6	1.7	9.23E-04
CTCCAG	P-loop containing nucleoside triphosphate hydrolase	107	7.5	1.5	0.001
CTCCAG	Intracellular	45	3.1	1.7	0.001
CTCCAG	Transferase	237	16.5	1.2	0.001
CTCCAG	Carbon metabolism	69	4.8	1.4	0.005
CTCCAG	Cytoplasm	151	10.5	1.3	0.007
CTCCAG	Cytosol	239	16.6	1.2	0.008
CTCCAG	Glyoxylate and dicarboxylate metabolism	30	2.1	1.7	0.015
CTCCAG	Fatty acid / polyketide synthesis	21	1.5	2.1	0.017
CTCCAG	Ligase	65	4.5	1.5	0.024
CACGCAG	Fatty acid / polyketide synthesis	18	2.5	3.5	9.26E-05
CACGCAG	Nucleotide-binding	90	12.5	1.4	0.02
CACGCAG	Cytosol	128	17.7	1.3	0.048
GATNNNNRTAC	Cell membrane	59	17.9	1.6	0.021
GATNNNNRTAC	Plasma membrane	149	45.3	1.2	0.023

Motifs were assigned to genes by finding overlap with coding regions. If found in intergenic regions the motif was assigned to the gene with the closest promoter. Genes at which the motif was found in >60% of the isolates were used to look for enrichment of pathways; * Bonferroni corrected P-value (P* in main text).

Supplementary Figure 1

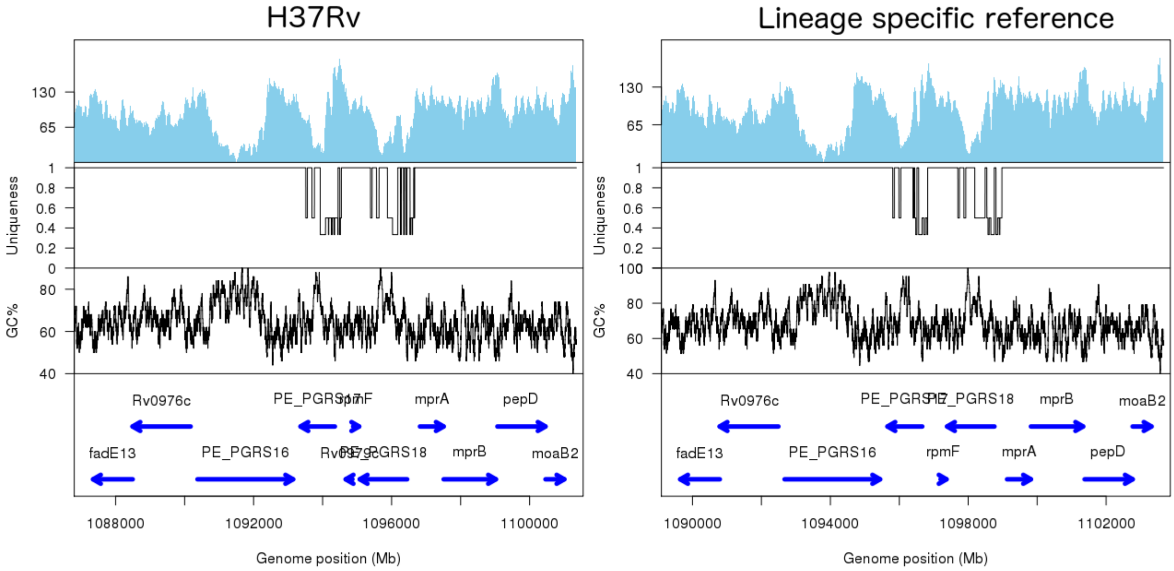
Differences in coverage in the *PE_PGRS3/4* and *PE_PGRS17/18* highly variable regions when comparing mapping of the WBB445_ARS7496 Illumina reads to the H37Rv reference and WBB445_ARS7496 Beijing reference described in this publication. The genes on the H37Rv reference used can be seen on the bottom track. The GC content and the uniqueness (1 = unique, < 1 non-unique) of a region can influence the coverage across the region and are plotted on the middle panels. The coverage is plotted on the top panel. The H37Rv mapping results are plotted on the left, while the WBB445_ARS7496 assembly results are plotted on the right.

A) *PE_PGRS3* region



Higher coverage is seen across both the *PE_PGRS3* and *PE_PGRS4* when mapping to the new lineage specific reference. Additionally, two new open reading frames have been introduced between the two genes.

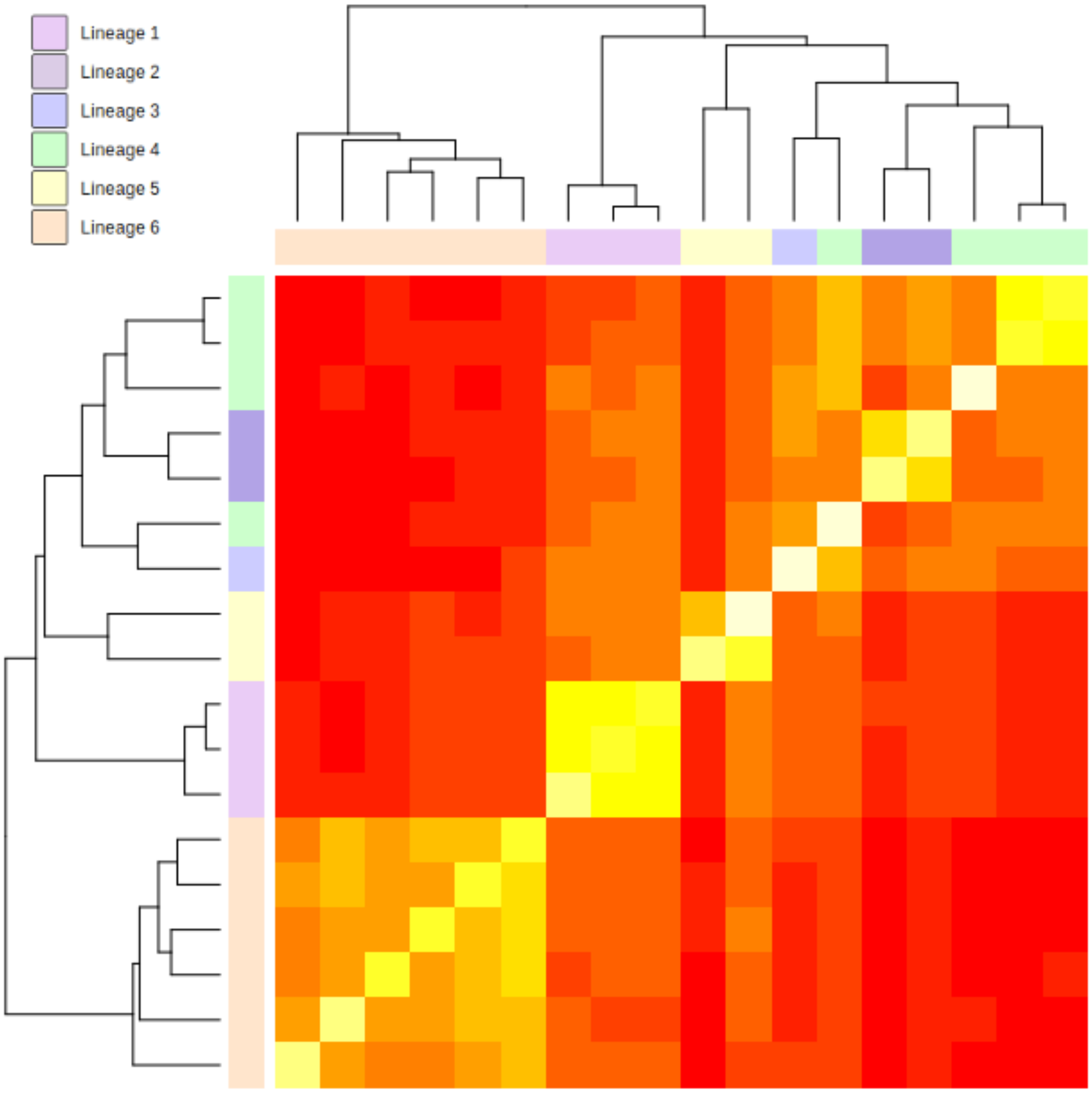
B) PE_PGRS17/18



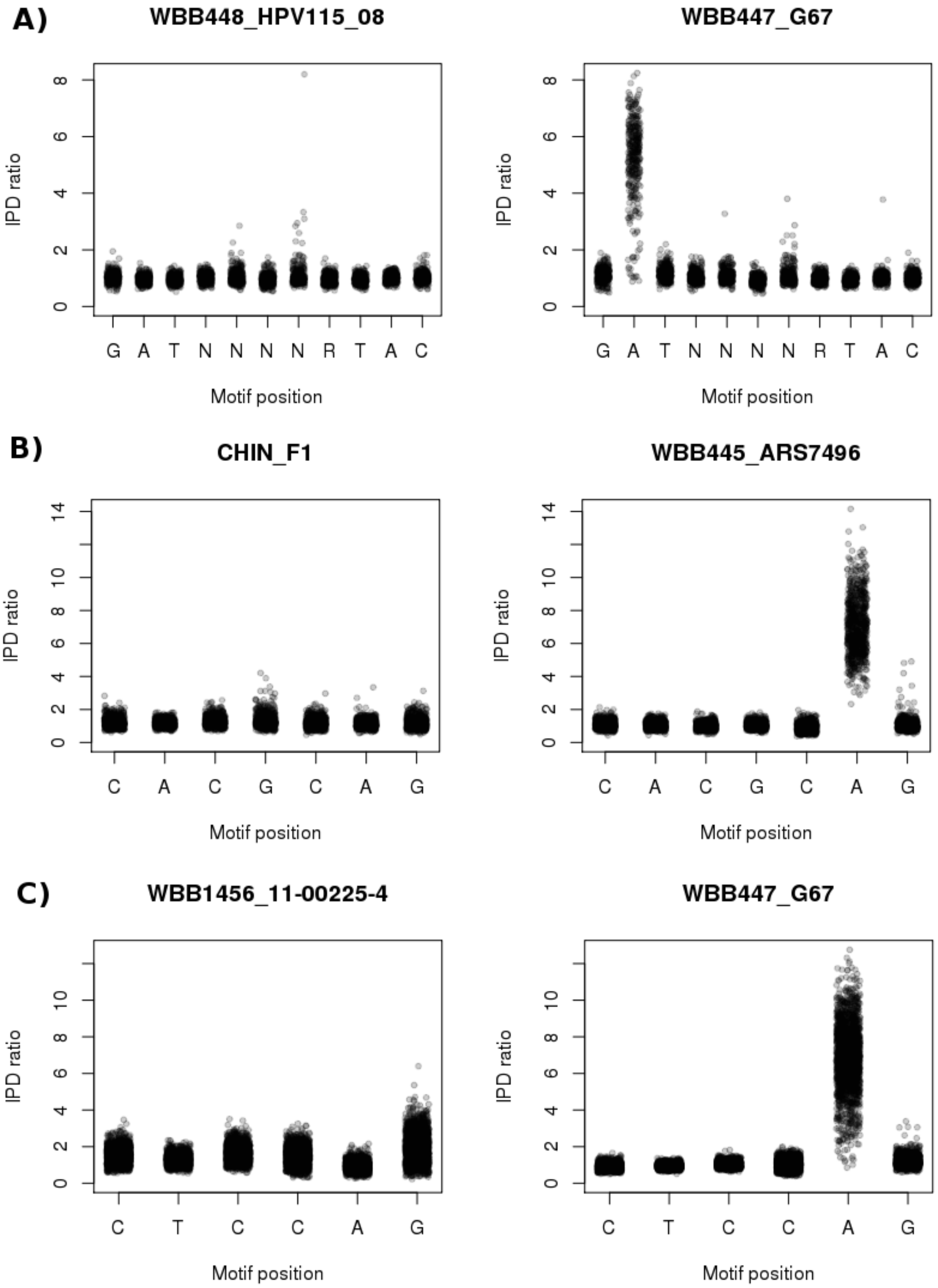
Only slight changes in genomic coverage were detected, indicating that the lack of coverage across these genes is mostly due to the high GC content in some regions coupled with the fact that some regions are non-unique.

Supplementary figure 2
Hierarchical clustering and heat map visualisation of shared number of orthogroups (groups of orthologous proteins).

Correct clustering can be observed for all isolates except CHIN_F1 (H37Rv strain) which is located outside lineage 4 and closer to lineage 3.

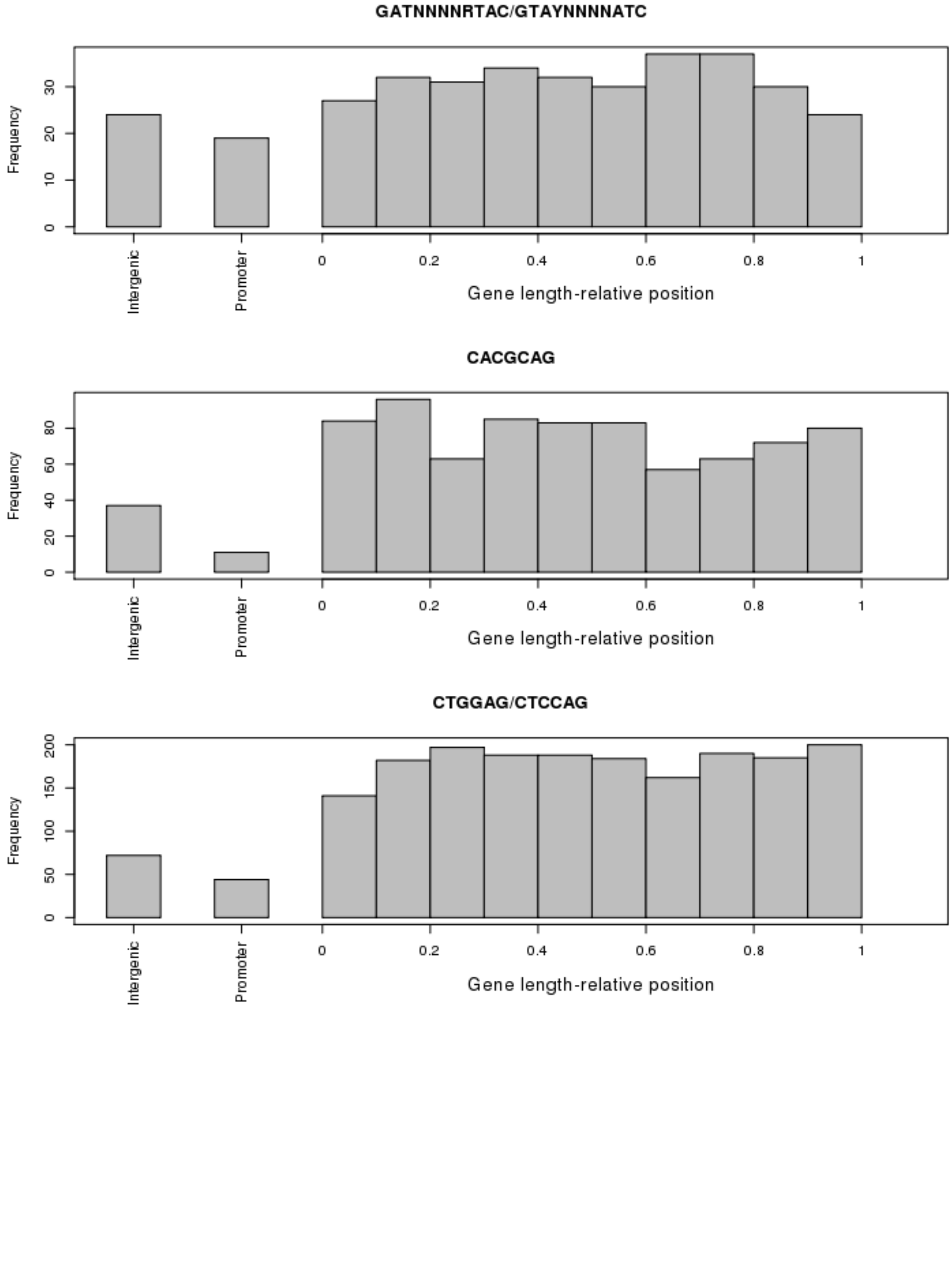


Supplementary figure 3
Inter-pulse duration (IPD) ratios across motifs in unmethylated isolates (left column) and methylated isolates (right column): A) GATN4RTAC, B) CACGCAG and C) CTCCAG



Supplementary figure 4

A histogram showing the location of the motifs relative to their associated genes. This plot was drawn for CHIN_F1, the H37Rv strain, and near identical distributions were seen for the other isolates. Where a motif is found in a coding region, its position relative to the gene length is shown. Most of the motifs are scattered randomly throughout the gene lengths and fewer are seen in the promoter.



RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Jody Phelan
Principal Supervisor	Taane Clark
Thesis Title	A Bioinformatic analysis of M. tuberculosis and host genomic data

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Scientific Reports
Please list the paper's authors in the intended authorship order:	Jody Phelan, Paola Florez de Sessions, Julian Parkhill, Surakameth Mahasirimongkol, Martin L. Hibberd, Taane G Clark
Stage of publication	**Not yet submitted**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I received raw sequence data for the *Mtb* isolates from our collaborators and subsequently proceeded with assessing the quality of the data. After trimming low quality data, I performed mapping and variants calling. Simultaneously, I processed the human dataset by performing imputation using a reference panel from the 1000 genomes project. After imputation, I removed all low frequency variants. I designed a script using a combination of C, python and R to establish a pipeline to analyse all dataset regarding this project. I also performed several iterations on the pipeline script to optimise all parameters. I curated the final results set and generated all plots and tables using R. I co-wrote the first draft of the manuscript.

Student Signature: ______________________________ **Date:** ______________

Supervisor Signature: ______________________________ **Date:** ______________

Chapter 8

Genome-wide host-pathogen analyses reveals genetic interaction points in tuberculosis disease

Genome-wide host-pathogen analyses reveals genetic interaction points in tuberculosis disease

Jody Phelan[1], Paola Florez de Sessions[2], Julian Parkhill [3], Surakameth Mahasirimongkol[4], Martin L. Hibberd[1,2,*], Taane G Clark[1,5,*]

[1] Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom

[2] Genomics Institute Singapore, Singapore

[3] Wellcome Trust Sanger Institute

[4] Medical Genetics Center, Medical Life Sciences Institute, Department of Medical Sciences, Ministry of Public Health, Nonthaburi, Thailand.

[5] Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom

* joint corresponding authors

ABSTRACT

Tuberculosis (TB) represents a major global health issue with an estimated 10.4 million new cases in 2015 alone. Innate susceptibility to tuberculosis has been a major focus of research in recent years. Genome wide association studies (GWAS) have been successfully applied to find loci associated with many infectious diseases including HIV, leprosy and Hepatitis C. This approach has not been fruitful for TB however, with lack of replication across study sites. The causal agent for TB, *M. tuberculosis* (*Mtb*), can be classified into seven distinct lineages which are differentially distributed geographically. The difference in locally circulating strains has been proposed as a reason for the lack of replication of GWAS hits. Here, we show that lineages and sub-lineages of TB are associated with specific variants in the human genome. We performed a genome-to-genome association using sequence data from the host and pathogen from 720 patients with pulmonary TB from Thailand. By performing association tests for each combination of variant found in both genomes we report a number of highly significant hits, including regions of the MHC. Markers for lineage one were highly associated with variants in the MHC region (rs2535298, $p=1.92\times10^{-10}$). Additionally, we found a number of sub-lineages and homoplastic variants in TB associated with loci in the human genome. The top hit ($p=5.36\times10^{-16}$) was between the SNP rs12548085 on chromosome 8p22 in the SGCZ gene. In total, thirty eight loci were highly associated ($\text{threshold}=1\times10^{-10}$) with specific pathogen variants. So far, GWASs have not considered the variation of the pathogen to be important for susceptibility. We present evidence of specific associations between human and could represent potential host-pathogen interactions.

INTRODUCTION

Tuberculosis disease (TB), caused by *Mycobacterium tuberculosis* is an important global public health issue, and high HIV prevalence and multi- and extensive-drug resistance are serious challenges to effective control[1]**.** There is an urgent need for better treatments and vaccines, which in turn require a deeper understanding of the biology of TB, which can be revealed by looking at the host-pathogen interaction check points that are exposed in the genetic signatures of human and *M. tuberculosis* genomes. Novel therapeutic approaches could be developed to exploit these interactions, including host immune modulators that mimic the successful natural responses seen in the majority of infections. It is possible to implement this type of approach within clinical trials and thereby attempt to modulate the human immune response to treat TB. However, without clear data on which strategies are successful in nature, these candidate approaches are unlikely to succeed.

M. tuberculosis genetics have been used extensively to describe its diversity. Sequence-based studies have characterised *M. tuberculosis* genomic variation, including single nucleotide polymorphisms (SNPs) and other variations such as insertions and deletions (indels), across thousands of samples[2,3]**.** Markers of drug resistance have been identified using phylogenetic tree-based and GWAS approaches[4]**.** Libraries of informative resistance mutations are leading to the development of informatic tools to rapidly profile samples for their drug susceptibility[5]**.** *M. tuberculosis* genetic regions under selective pressure, perhaps due to drug resistance or host immune responses, can be detected[4,6]**.** *M. tuberculosis* has seven lineages that are endemic in different locations around the globe, leading to the hypothesis that the strain-types are specifically adapted to people of different genetic backgrounds[7]**.** The lineages vary in their geographic distribution and spread, with lineage 2 being particularly mobile with evidence of recent spread from Asia to Europe and Africa[8]**.** Lineage 4 is common in Europe and southern Africa, with regions of high TB incidence and high levels of HIV co-infection. The lineages may vary in propensity to transmit, to cause disease, in the site and severity of disease[9–11]**,** but results are inconsistent and there is considerable inter-strain variation within lineages[12,13]**.** A set of SNPs has been identified that can be used to barcode sub-lineages[2]**,** leading to informatic tools that position sequenced samples within a global phylogeny[14]. Similarly, SNPs have been used to construct transmission networks, where samples from different individuals that have near identical genomic variation are most likely to be due to a transmission event. Inferred transmission chains based on genome-wide SNPs in northern Malawian isolates has shown striking differences by lineage in the proportion of disease due to recent transmission and in transmissibility (highest- lineage-2 (East-Asian), lowest- lineage-1 (Indo-Oceanic)) that were not confounded by HIV status or drug resistance[15–17]**.**

Host genetics has the potential to inform about TB disease susceptibility. Despite the GWAS successes in the infectious disease field[18–21], this approach has proven difficult for TB[22–24], with the susceptibility loci identified not replicated across studies[25,26]. Reasons for non-replication include differences in human population structure or variation in *M. tuberculosis* strains, but also in TB case definitions; and controls being a mixture of unexposed, exposed, and latently infected individuals. Despite the general difficulties of TB GWAS approaches to date, promising recent work has shown that the human leukocyte antigen (HLA) class II region contributes to genetic risk of pulmonary

TB, described as possibly acting through reduced presentation of protective *M. tuberculosis* antigens to T cells[27]. Recent work in TB meningitis has revealed a different potential susceptibility pathway, which needs to be replicated (**Hibberd et al, in preparation**). However, the GWAS approach applied to another mycobacterial infection, leprosy, has identified the innate intracellular signalling pathways involved in macrophage killing of bacteria (the NOD pathway and RAB32) as critical to leprosy outcome[28,29], and also linked to Crohn's disease[30], suggesting that the approach could be successful. To date there have been no robust studies of human-*M. tubercul*osis interaction genomics.

Host-pathogen interaction genomics has already begun to be used to identify pathogenic mechanisms associated with other diseases, including meningococcal disease[31] and hepatitis C virus infection[32]. There is some evidence to suggest host-pathogen effects impact on *M. tuberculosis,* and it is able to subvert the human host response to infection, including the persistent nature of the infection and the possibility of multiple re-infections; although the mechanisms of this process remain unclear. The *M. tuberculosis pe/p*pe gene families (~10% of genome) are hypothesised to include important virulence factors involved with host-pathogen interactions[33]. There is evidence of innate and adaptive human host responses to *M. tuberculosis*, with B and T cell recognition of pe/ppe proteins[34]. These proteins may represent a source of antigenic variation, which allow the organism to escape antigen-specific host responses[34]. With *M. tuberculosis* antigens being presented through HLA molecules, there is a strong argument for assessing the interaction between *M. tuberculosis pe/ppe* and human HLA genotypes[35,36]. However, because *pe/ppe* genes are highly variable and complex to analyse, they are typically disregarded in genome studies. Recent SNP analysis has revealed that variation in the majority of the 168 *pe/ppe* genes studied is consistent with *M. tuberculo*sis lineage[33]. Evidence of positive selection was revealed in 65 *pe/ppe* genes, including epitopes potentially binding to major histocompatibility complex (MHC) molecules.

By integrating the human and *M. tuberculosis* genetics data in a well characterised Thailand cohort (n= 720), we sought to reveal insights into interaction points using "genome-to-genome" analytical methods. Our analyses reveal a crucial role for the HLA, and previously unknown genes such as *CNTN3* and *USP6NL*.

RESULTS

M. tuberculosis genetic diversity in the Thai cohort

Host-*M. tuberculosis* genetic data were complete for 720 TB cases. *In silico* profiling using TB Profiler[5] determined that isolates were predominantly from lineages 1 (35%), 2 (47%) and 4 (16%) (lineage 3 <2%) (**Supplementary table 1**). The isolates were predicted to be predominantly pan-susceptible across 14 drugs (96.2%) with the remainder being multidrug resistant (isoniazid and rifampicin, 3.8%), and none extensively-drug resistant. Raw reads were trimmed and mapped to the H37Rv reference genome (AL123456), and 59k high quality unique variants were called. The vast majority (95.2%) were rare variants with minor allele frequencies less than 5%. Phylogenetic reconstruction and principal component analysis (PCA) revealed a strong population stratification with strong clustering by lineage (**Figure 1a**).

Human genetic diversity in the Thai cohort

Human genotypes were imputed using Asian populations from the 1000 Genomes Project phase 3, resulting in ~6 million high quality variants with minor allele frequency > 5%. Using these SNPs in a PCA approach, the individuals clustered into three groups that coincide with Thai ethnic diversity (**Figure 1b, c**). The proportion of each lineage within each group was calculated, and revealed an unequal distribution of lineage 1 strains between groups (**Supplementary table 1**).

Genome-to-lineage analysis

To investigate differential susceptibility to *M. tuberculosis* lineages, within a regression framework we tested for associations between human variants as predictors and lineages as the outcome variable. Each lineage was compared against all other lineages in a case-control type analysis. We did not consider lineage 3 in this analysis as the sample size is only eight. At an established significance cut-off (1×10^{-8}), we identified putative associations for lineage 1 (66 SNPs, 13 loci) and lineage 4 (7 SNPs, 4 loci), but not lineage 2 (**Table 1**). For lineage 1, the most significant association was found to be shared between three SNPs (two in *C6orf15*, one in a pseudogene) located within the MHC class I region (**Supplementary Figure 1**). For lineage 4, the strongest association was the present in the *USP6NL* gene (variant: rs4750068). To follow-up the HLA-Lineage1 association, we imputed HLA haplotypes using SNP2HLA software, and re-tested for association to lineage 1. Though no haplotype reached the 10^{-8} cut-off, the most significant association was present at the *DQA1* locus (type: 06:01, p-value = 8.9e-8) (**Supplementary table 2**).

Genome-to-genome analysis reveals host-pathogen interactions

There is sequence diversity within lineages and the previous approach may miss potential interactions between human variants and sub-lineages or homoplastic *M. tuberculosis* variants. To identify these potential interactions, we applied a regression-based approach using *M. tuberculosis* alleles as phenotypes and testing for epistatic effects between the ~6M human and 2,002 *M. tuberculosis* SNPs (MAF >5%). At an established significance cut-off (1 x 10^{-10}), this approach revealed associations involving 199 human SNPs (38 loci) (**Figure 2**) (**Table 1**). Associations to lineage, sub-lineage, and homoplastic SNPs were found. The strongest association signal was between the rs12548085 SNP (*SGCZ* gene) and a subclade in lineage 1 (p=5.36×10^{-16}). Other noteworthy genes found, include: *HDAC4* (p=2.06×10^{-12}, lineage 1.1) and *PRKCA* (p=4.88×10^{-11}, lineage 4.5) and *TNFSF9* (p=4.86×10^{-11}, lineage 2.2.1). A homoplastic SNP in the *Rv3467* gene (K315E) was associated with the human polymorphism rs9398635 (chr. 6, intergenic region) p=5.63×10^{-14}).

DISCUSSION

There have been a number of attempts to identify loci that influence susceptibility to tuberculosis[22,24]. While statistically significant loci have been reported, they have not been validated in across populations[23,26]. It has been postulated that *M. tuberculosis* has been in a state of co-evolution with its host[37], and by implication there are differences in human population susceptibility to infections from different lineages. This observation could explain the lack of reproducibility of hits found in the different GWAS. To detect whether there are any human variants influencing the likelihood of infection of a particular *M. tuberculosis* lineage or sub-lineage, we performed a genome-to-genome analysis using a GWAS approach. Association patterns within the *M. tuberculosis* genome reflected intra- lineage or sub-lineage-specific, or inter-lineage (homoplastic) effects.

For intra-effects, a single human polymorphism will be associated to many *M. tuberculosis* variants with equal statistical significance due to the clonal nature (lack of recombination) leading to high linkage disequilibrium, and long branches leading up to the lineages (**Figure 1a**). The resolution to which we can narrow down the list of possible causal variants depends on the sampling depth and the effect size of the allele. Phenotypic bacterial differences are known to result from strain- and lineage- specific variation. Although efforts have mainly focused on the phenotypic differences in drug resistance, transmissibility and virulence[16,38,39], it cannot be ruled out that this variation also contributes towards host susceptibility. For inter-lineage effects, homoplastic variants appear throughout the phylogenetic tree and rarely share the same pattern of variation with other variants, therefore it is possible to localise to a specific *M. tuberculosis* variant that is driving the association.

To detect whole-lineage signals, we performed GWAS using the *M. tuberculosis* lineages as the phenotype. The most significant P-value occurred between markers at the MHC locus and lineage 1. This indicates that one or more variants acquired after the divergence of the “ancient” and “modern” lineages have a significant association with variants in the MHC class I region. The Manhattan plot for the entire MHC region reveals another peak at the MHC class II region, though this does not reach the significance cut-off. An analysis of imputed HLA haplotypes points to the HLA DQA1*06:01 type to have the highest significance, and further supports other studies implicating variation of MHC region in tuberculosis susceptibility[27,35,36,40], although this interaction effect may be strain-specific. Consequently, human populations could differ in their susceptibility to different lineages of *M. tuberculosis* and this finding supports the host-pathogen coevolution hypothesis. Twelve additional loci were associated with lineage 1 including a variant in the *CNTN3* gene (rs34989253). *CNTN3* and bovine MHC complex have been previously implicated in susceptibility to Bovine leukaemia virus[41]. Four human loci were associated with lineage 4, but no previous associations of these genes to infectious disease could be found.

To uncover intra and inter-lineage or convergent evolution variants interacting with human polymorphisms, we undertook a more agnostic approach and performed a GWAS using the *M. tuberculosis* alleles as phenotypes. While we retain the MHC-lineage 1 association, we found many additional low-frequency variants within 38 loci. One putative association involved rs7251888 in *TNFSF9* and a subclade of lineage 2.2.1. *TNFSF9* is a cytokine involved with antigen presentation in T cells and has been proposed as a useful marker in the detection of *M. tuberculosis*-reactive CD4$^+$ T cells[42]. It has also been proposed to regulate innate and adaptive immpune response against *Mtb*[43,44]. Variants in *HDAC4* and *PRKCA,* which are both involved in response to interleukins[45,46], were associated with subclades of lineage 1.1 and lineage 4.5 respectively. Several homoplastic variants were also associated with human variants, the most significant between an intergenic SNP on chromosome 6, close to the *GJA1* gene and the *M. tuberculosis Rv3467* (K315E) SNP. Other significant homoplastic variants in *M.*

tuberculosis included *ppe18* and *mceF*, which have implicated roles in intracellular survival[47,48] (**Table 2**).

This study has highlighted the importance of the MHC region in susceptibility to tuberculosis and specific strain-types, implying it is a crucial interaction point. Interestingly, many of the new hits discovered using the genome-to-genome approach have a much more significant P-value, with the minimum reaching 5.36×10. While performing association on a lineage highlights regions of interest, it may not be enough. A considerable amount of variation exists within lineages and within populations to which they are endemic to. By testing all possible combinations of variants we have highlighted many significant associated variants. This suggests that susceptibility to tuberculosis follows a complicated pattern with many host factors involved coupled with the diversity within the *M. tuberculosis* pathogen. The relative importance of these interactions must be investigated through follow up studies in different populations.

ONLINE METHODS

Study population

The Thailand cases (HIV negative TB patients with no known previous TB (age > 14 years)) were from Chiang Rai, Lampang and Bangkok provinces (TB incidence 181/100,000 population).

Genetic data

Human genotypes for the Thai TB cases (n=720) were generated on Illumina Human610-Quad BeadChip and Illumina HumanOmniExpressExome-8 v1.2 BeadChip, complemented by imputation of >8.4 million genomic sites using BEAGLE4.1 software[49] and a 1000 Genomes reference panel[50]. HLA protein alleles were imputed using SNP2HLA software and a pan-Asian reference[51]. SNPs were removed if there was: (i) deviation in genotypic frequencies from Hardy-Weinberg equilibrium (HWE) as assessed using a chi-square test (P<0.00001); (ii) high genotype call missingness (>10%); (iii) low minor allele frequency (<5%); or (iiii) low imputation quality (allelic R^2<0.7). The population structure was explored using principal component analysis inferred from pairwise SNP genotype differences between individuals.

Pathogen sequence data was generated at the Sanger institute using an Illumina HiSeq 2000 machine. Raw *M. tuberculosis* sequencing data was aligned to the H37Rv reference genome (Genbank accession number: NC_000962.3) using the *BWA mem* algorithm[52]**.** The SAMtools/BCFtools[53] software was used to call SNPs and small indels using default options. Alleles were additionally called across the whole genome (including SNP sites) using a coverage-based approach. A missing call was assigned if the total depth of coverage at a site did not reach a minimum of 20 reads or none of the four nucleotides accounted for at least 75% of the total coverage. Samples or SNP sites having an excess of 10% missing genotype calls were removed. This quality control step was implemented to remove samples with bad quality genotype calls due to poor depth of coverage or mixed infections. The final discovery dataset included 720 Thai isolates and ~59k genome-wide SNPs. Lineages were predicted using the TBProfiler tool[5]. The analytical pipeline is described in greater detail elsewhere[3,4].

Statistical analysis

To uncover effects between lineages and human genotypes, a separate logistic region model was fitted for each lineage (lineage X vs. lineage non-X) using the plink 1.9[54] software, with a full model analysis (--model) setting. The minimum p-value across the tests for each variant was retained. A statistical significance threshold was established by simulations ($P < 1 \times 10^{-8}$). The genome-to-genome analysis was performed using the same modelling strategy, except we used the *M. tuberculosis* alleles (minor vs. major) as the outcome. *M. tuberculosis* variants were included in the analysis if (i) they were not synonymous; (ii) had a minor allele frequency >0.05 and (iii) were not solely located in transmission clusters (median SNP distance between isolates with mutation >20). A statistical significance threshold was established by simulation ($P < 1 \times 10^{-10}$). Regional association plots were generated using locuszoom[55].

DATA AVAILABILITY

GWAS genotypic data is shared through the EBI European Genome-phenome initiative. All pathogen raw sequencing data is available from PRJEB7056.

ACKNOWLEDGMENTS

JPh is funded by a BBSRC PhD studentship. TGC is funded by the Medical Research Council UK (Grant no. MR/K000551/1 and MR/M01360X/1, MR/N010469/1, MC_PC_15103). SC is funded by Medical Research Council UK grants (MR/M01360X/1, MC_PC_15103). We gratefully acknowledge the Scientific Computing Group for data management and compute infrastructure at Genome Institute of Singapore for their help. The MRC eMedLab computing resource was used for bioinformatics and statistical analysis. The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

MLH and TGC conceived and directed the project. SM led the sample collection and DNA extraction. PFdS and MLH coordinated genotyping. JPa coordinated sequencing. JPh performed bioinformatic and statistical analyses under the supervision of MLH and TGC. JPh, PFdS, MLH and TGC interpreted results. JPh, MLH and TGC wrote the first draft of the manuscript. All authors commented and edited on various versions of the draft manuscript and approved the final manuscript. JPh, MLH and TGC compiled the final manuscript.

DISCLOSURE DECLARATION

There are no conflicts of interest.

REFERENCES

1. Organisation, W. H. Global tuberculosis report 2016. (2016).
2. Coll, F. *et al.* A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nat. Commun.* **5,** 4812 (2014).
3. Coll, F. *et al.* PolyTB: A genomic variation map for Mycobacterium tuberculosis. *Tuberculosis* **94,** 346–354 (2014).
4. Phelan, J. *et al.* Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC Med.* **14,** 31 (2016).
5. Coll, F. *et al.* Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med.* **In Press,** (2015).
6. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. *Am. J. Hum. Genet.* **90,** 7–24 (2012).
7. Reed, M. B. *et al.* Major Mycobacterium tuberculosis lineages associate with patient country of origin. *J. Clin. Microbiol.* **47,** 1119–28 (2009).
8. European Concerted Action on New Generation Genetic Markers and Techniques for the Epidemiology and Control of Tuberculosis. Beijing/W Genotype *Mycobacterium tuberculosis* and Drug Resistance. *Emerg. Infect. Dis.* **12,** 736–743 (2006).
9. Click, E. S., Moonan, P. K., Winston, C. A., Cowan, L. S. & Oeltmann, J. E. Relationship Between Mycobacterium tuberculosis Phylogenetic Lineage and Clinical Site of Tuberculosis. *Clin. Infect. Dis.* **54,** 211–219 (2012).
10. Krishnan, N. *et al.* Mycobacterium tuberculosis Lineage Influences Innate Immune Response and Virulence and Is Associated with Distinct Cell Envelope Lipid Profiles. *PLoS One* **6,** e23870 (2011).
11. Gagneux, S. Host-pathogen coevolution in human tuberculosis. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367,** 850–9 (2012).
12. Portevin, D., Gagneux, S., Comas, I., Young, D. & Belardelli, F. Human Macrophage Responses to Clinical Isolates from the Mycobacterium tuberculosis Complex Discriminate between Ancient and Modern Lineages. *PLoS Pathog.* **7,** e1001307 (2011).
13. Mathema, B. *et al.* Epidemiologic Consequences of Microvariation in Mycobacterium tuberculosis. *J. Infect. Dis.* **205,** 964–974 (2012).
14. Benavente, E. D. *et al.* PhyTB: Phylogenetic tree visualisation and sample positioning for M. tuberculosis. *BMC Bioinformatics* **16,** 155 (2015).
15. Glynn, J. R. *et al.* Whole Genome Sequencing Shows a Low Proportion of Tuberculosis Disease Is Attributable to Known Close Contacts in Rural Malawi. *PLoS One* **10,** e0132840 (2015).
16. Guerra-Assunção, J. *et al.* Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *Elife* **4,** (2015).
17. Guerra-Assunção, J. A. *et al.* Recurrence due to Relapse or Reinfection With *Mycobacterium tuberculosis* : A Whole-Genome Sequencing Approach in a Large, Population-Based Cohort With a High HIV Infection Prevalence and Active Follow-up. *J. Infect. Dis.* **211,** 1154–1163 (2015).
18. Khor, C. C. & Hibberd, M. L. Host–pathogen interactions revealed by human genome-wide surveys. *Trends Genet.* **28,** 233–243 (2012).

19. Khor, C.-C. & Hibberd, M. L. Revealing the molecular signatures of host-pathogen interactions. *Genome Biol.* **12,** 229 (2011).
20. Khor, C. C. & Hibberd, M. L. Shared pathways to infectious disease susceptibility? *Genome Med.* **2,** 52 (2010).
21. Jallow, M. *et al.* Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nat. Genet.* **41,** 657–665 (2009).
22. Curtis, J. *et al.* Susceptibility to tuberculosis is associated with variants in the ASAP1 gene encoding a regulator of dendritic cell migration. *Nat. Genet.* **47,** 523–527 (2015).
23. Hu, X. *et al.* No Significant Effect of ASAP1 Gene Variants on the Susceptibility to Tuberculosis in Chinese Population. *Medicine (Baltimore).* **95,** e3703 (2016).
24. Thye, T. *et al.* Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. *Nat. Genet.* **42,** 739–741 (2010).
25. Png, E. *et al.* A genome wide association study of pulmonary tuberculosis susceptibility in Indonesians. *BMC Med. Genet.* **13,** 5 (2012).
26. Mahasirimongkol, S. *et al.* Genome-wide association studies of tuberculosis in Asians identify distinct at-risk locus for young tuberculosis. *J. Hum. Genet.* **57,** 363–367 (2012).
27. Sveinbjornsson, G. *et al.* HLA class II sequence variants influence tuberculosis risk in populations of European ancestry. *Nat. Genet.* **48,** 318–322 (2016).
28. Zhang, F.-R. *et al.* Genomewide Association Study of Leprosy. *N. Engl. J. Med.* **361,** 2609–2618 (2009).
29. Zhang, F. *et al.* Identification of two new loci at IL23R and RAB32 that influence susceptibility to leprosy. *Nat. Genet.* **43,** 1247–1251 (2011).
30. Schurr, E. & Gros, P. A Common Genetic Fingerprint in Leprosy and Crohn's Disease? *N. Engl. J. Med.* **361,** 2666–2668 (2009).
31. Davila, S. *et al.* Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease. *Nat. Genet.* **42,** 772–776 (2010).
32. Ansari, M. A. *et al.* Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis C virus. *Nat. Genet.* **49,** 666–673 (2017).
33. Phelan, J. E. *et al.* Recombination in pe/ppe genes contributes to genetic variation in Mycobacterium tuberculosis lineages. *BMC Genomics* **17,** 151 (2016).
34. Sampson, S. L. Mycobacterial PE/PPE Proteins at the Host-Pathogen Interface. *Clin. Dev. Immunol.* **2011,** 1–11 (2011).
35. Ehlers, S. & Schaible, U. E. The granuloma in tuberculosis: dynamics of a host-pathogen collusion. *Front. Immunol.* **3,** 411 (2012).
36. Brodin, P. *et al.* High Content Phenotypic Cell-Based Visual Screen Identifies Mycobacterium tuberculosis Acyltrehalose-Containing Glycolipids Involved in Phagosome Remodeling. *PLoS Pathog.* **6,** e1001100 (2010).
37. Gagneux, S. Host-pathogen coevolution in human tuberculosis. *Philos. Trans. R. Soc. B Biol. Sci.* **367,** 850–859 (2012).
38. Cox, H. S. *et al.* The Beijing genotype and drug resistant tuberculosis in the Aral Sea region of Central Asia. *Respir. Res.* **6,** 134 (2005).
39. Dunn, P. L. & North, R. J. Virulence ranking of some Mycobacterium tuberculosis and Mycobacterium bovis strains according to their ability to multiply in the

lungs, induce lung pathology, and cause mortality in mice. *Infect. Immun.* **63,** 3428–37 (1995).

40. Duarte, R. *et al.* HLA class II alleles as markers of tuberculosis susceptibility and resistance. *Rev. Port. Pneumol.* **17,** 15–9

41. Takeshima, S., Sasaki, S., Meripet, P., Sugimoto, Y. & Aida, Y. Single nucleotide polymorphisms in the bovine MHC region of Japanese Black cattle are associated with bovine leukemia virus proviral load. *Retrovirology* **14,** 24 (2017).

42. Yan, Z. -h. *et al.* CD137 is a Useful Marker for Identifying CD4 $^+$ T Cell Responses to *Mycobacterium tuberculosis*. *Scand. J. Immunol.* **85,** 372–380 (2017).

43. Fernández Do Porto, D. A. *et al.* CD137 differentially regulates innate and adaptive immunity against Mycobacterium tuberculosis. *Immunol. Cell Biol.* **90,** 449–456 (2012).

44. Martínez Gómez, J. M. *et al.* Role of the CD137 ligand (CD137L) signaling pathway during Mycobacterium tuberculosis infection. *Immunobiology* **219,** 78–86 (2014).

45. Boudreau, R. T. M., Garduno, R. & Lin, T.-J. Protein Phosphatase 2A and Protein Kinase Cα Are Physically Associated and Are Involved in *Pseudomonas aeruginosa* -induced Interleukin 6 Production by Mast Cells. *J. Biol. Chem.* **277,** 5322–5329 (2002).

46. Han, S. *et al.* Recruitment of histone deacetylase 4 by transcription factors represses interleukin-5 transcription. *Biochem. J.* **400,** 439–48 (2006).

47. McCann, J. R., McDonough, J. A., Sullivan, J. T., Feltcher, M. E. & Braunstein, M. Genome-wide identification of Mycobacterium tuberculosis exported proteins with roles in intracellular growth. *J. Bacteriol.* **193,** 854–61 (2011).

48. Bhat, K. H., Ahmed, A., Kumar, S., Sharma, P. & Mukhopadhyay, S. Role of PPE18 protein in intracellular survival and pathogenicity of Mycobacterium tuberculosis in mice. *PLoS One* **7,** e52601 (2012).

49. Browning, B. L. & Browning, S. R. Genotype Imputation with Millions of Reference Samples. *Am. J. Hum. Genet.* **98,** 116–126 (2016).

50. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

51. Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One* **8,** e64683 (2013).

52. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).

53. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–2993 (2011).

54. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

55. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26,** 2336–2337 (2010).

Table 1
Genome-to-lineage association results

Lineage	Chr.	SNP ID	P-value	Gene ID	Distance*	OR
1	6	rs2535298	1.92E-10	*C6orf15*	14538	0.370
1	18	rs359758	3.13E-10	*RP11*	233049	2.672
1	9	rs10810134	8.20E-10	*NFIB*	0	0.262
1	20	rs6140144	9.52E-10	*SLC52A3*	0	1.985
1	2	rs4668246	1.28E-09	*MYO3B*	0	1.980
1	18	rs57810761	3.22E-09	*SERPINB8*	415380	2.564
1	4	rs10657819	3.41E-09	*CCSER1*	159196	0.489
1	14	rs1075612	3.58E-09	*FLRT2*	7984	2.103
1	2	rs113625848	5.23E-09	*CNTNAP5*	947438	1.979
1	10	rs7095852	5.70E-09	*PCBD1*	12435	0.513
1	13	rs7334180	5.84E-09	*FLT3*	0	2.008
1	11	rs10897830	6.17E-09	*FAM181B*	1228892	0.398
1	19	rs4024210	7.01E-09	*PDE4A*	1437	2.213
1	14	rs58579744	7.62E-09	*VRK1*	4500	1.801
1	21	rs73182460	7.65E-09	*APP*	0	2.167
1	3	rs34989253	7.98E-09	*CNTN3*	0	0.513
1	11	rs11320420	8.76E-09	*MYRF*	0	0.507
4	10	rs4750068	1.09E-09	*USP6NL*	0	1.921
4	17	rs138005149	1.54E-09	*AC102948.2*	29772	2.657
4	6	rs6924775	3.51E-09	*PRR18*	39544	3.977
4	18	rs142396797	3.69E-09	*FHOD3*	0	3.718
4	2	rs4854538	6.85E-09	*ANTXR1*	0	1.570

* Distance to closest annotated CDS

Table 2
Genome-to-genome association results

The minimum p-value per gene and the associated odds ratio and lineage of the *Mtb* variant * Distance to closest annotated CDS; ** Lineage associated with *Mtb* variant

Chr.	Human SNP	Host Gene	P-value	Distance*	OR	Lineage**	*Mtb* SNP
8	rs12548085	*SGCZ*	5.36E-16	0	2.542	lineage1.1.1	
4	rs7670123	*CLNK*	5.07E-15	288813	2.722	lineage2.2.1	
6	rs9398635	*GJA1*	5.63E-14	25506	0.304	lineage1, lineage2, lineage3, lineage4	*Rv3467 K315E*
17	rs6504803	*C17orf112*	1.11E-13	37174	3.045	lineage4.5	
7	rs1149213	*SEMA3E*	4.55E-13	46140	3.792	lineage4.5	
7	rs672365	*AGR3*	4.63E-13	178511	2.057	lineage1, lineage2, lineage4	*Rv0336 H496P*
8	rs75969446	*SOX17*	8.80E-13	21659	2.859	lineage4.5	
19	rs55916171	*UQCRFS1*	1.70E-12	1044300	3.052	lineage1.1.1	
2	rs291333	*HDAC4*	2.06E-12	0	3.348	lineage1, lineage2	*mce3F A170R*
5	rs59612284	*ADAMTS16*	2.58E-12	0	4.667	lineage4.5	
19	rs117476816	*SULT2B1*	6.89E-12	0	3.048	lineage4.5	
14	rs7144346	*RP11*	8.25E-12	4575	2.615	lineage1, lineage2, lineage4	*ppe18 S263H*
3	rs542038782	*SLITRK3*	9.94E-12	152879	3.822	lineage4.5	
20	rs6140144	*SLC52A3*	1.21E-11	0	2.131	lineage1	
13	rs7983548	*INTS6*	1.22E-11	0	2.937	lineage4.5	
13	rs145372612	*SERPINE3*	1.22E-11	0	2.982	lineage4.5	
6	rs2535298	*C6orf15*	2.12E-11	14538	0.346	lineage1	
4	rs6536724	*NPY5R*	2.63E-11	7713	3.431	lineage1	

Table 2 - continued

Chr.	Human SNP	Host Gene	P-value	Distance*	OR	Lineage**	*Mtb* SNP
5	rs6601202	*FAM153C*	3.10E-11	5257	3.051	lineage1, lineage2	*mce3F* *A170R*
16	rs900729	*ANKRD11*	4.11E-11	0	0.328	lineage4	
8	rs7013247	*SNTG1*	4.37E-11	0	2.8	lineage4.5	
17	rs2525103	*COPZ2*	4.37E-11	15974	2.625	lineage4.5	
10	rs116986894	*GATA3*	4.42E-11	513865	3.228	lineage1.1.1	
13	rs17075761	*WDFY2*	4.43E-11	52822	3.486	lineage4.5	
12	rs7133564	*MUCL1*	4.58E-11	0	2.399	lineage1.1	
22	rs9605254	*CECR2*	4.71E-11	40429	13.94	lineage1, lineage2, lineage4	*PE_PGRS56* *N679A*
2	rs17026212	*TGOLN2*	4.86E-11	0	2.464	lineage1.1.1	
6	rs116672827	*NEDD9*	4.86E-11	0	2.045	lineage1.1.1	
19	rs7251888	*TNFSF9*	4.86E-11	8847	2.391	lineage2.2.1	
17	rs77462363	*PRKCA*	4.88E-11	0	5.95	lineage4.5	
6	rs9322189	*GINM1*	5.60E-11	0	2.098	lineage2.2.1	
18	rs359758	*RP11*	5.62E-11	233049	2.789	lineage1	
8	rs10097239	*TUSC3*	6.65E-11	140682	3.303	lineage4.5	
1	rs6675820	*DNTTIP2*	8.26E-11	0	2.346	lineage1, lineage2, lineage4	*ppe18* *S263H*
7	rs74918833	*ISPD*	8.27E-11	38372	3.239	lineage4.5	
15	rs2467365	*C15orf41*	8.80E-11	113668	2.368	lineage1.1	
6	rs34607745	*HSF2*	9.26E-11	40356	3.917	lineage1.1.1	
2	rs34312950	*REG3G*	9.53E-11	126939	2.505	lineage1, lineage2, lineage4	*ppe18* *S263H*

Figure 1

Principal component analysis (PCA) of *M. tuberculosis* and human genotypes

(a) Phylogenetic tree of the *M. tuberculosis* in Thailand; (b) PCA of the human variants was performed followed by k-means clustering, leading to three main clusters; (c) The lineages associated with each patient was then visualised with the clusters superimposed. A noticeable difference in the number of lineage 1 strains was evident (see Supplementary Table 1); (d) Genome-to-genome interactions revealed

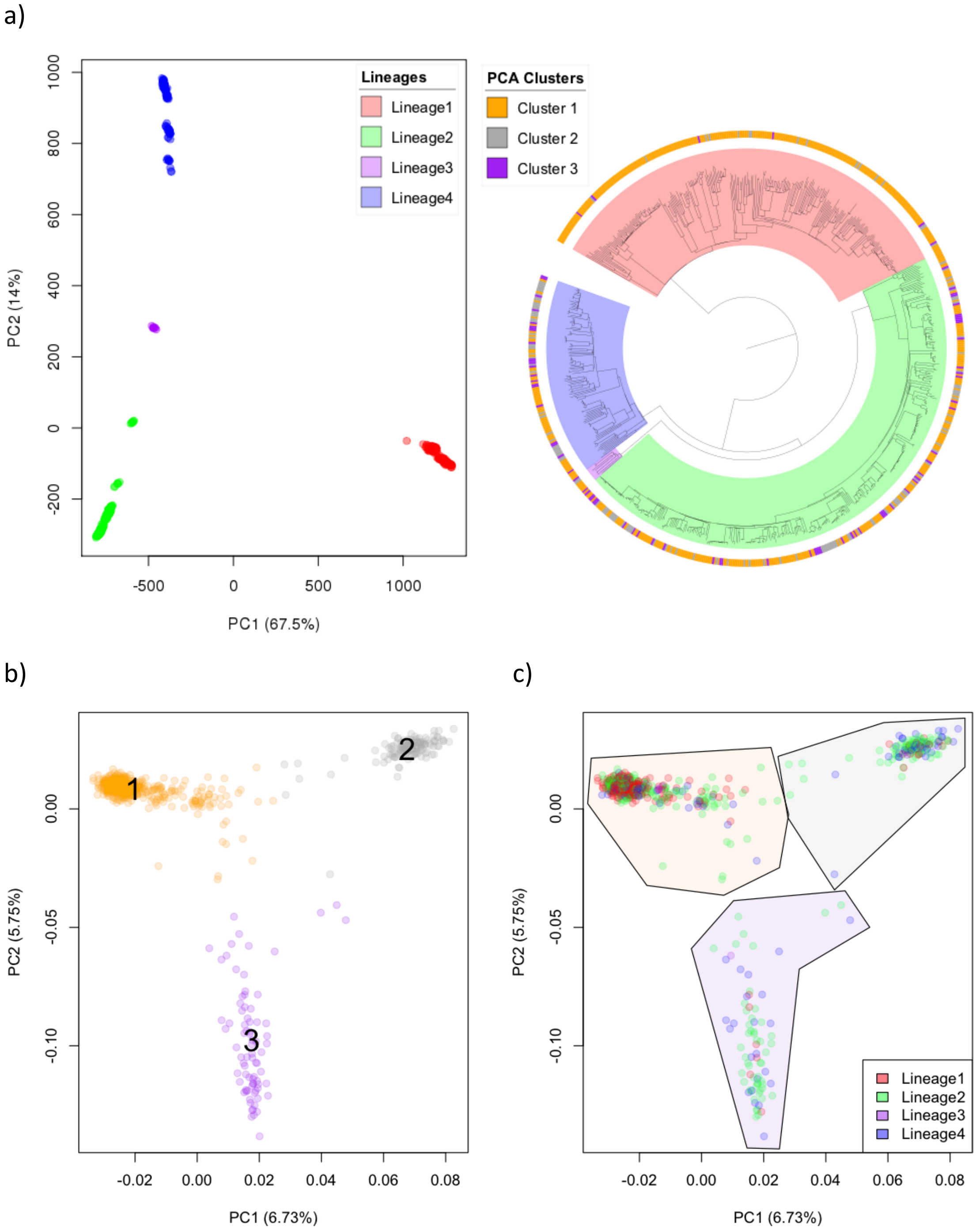
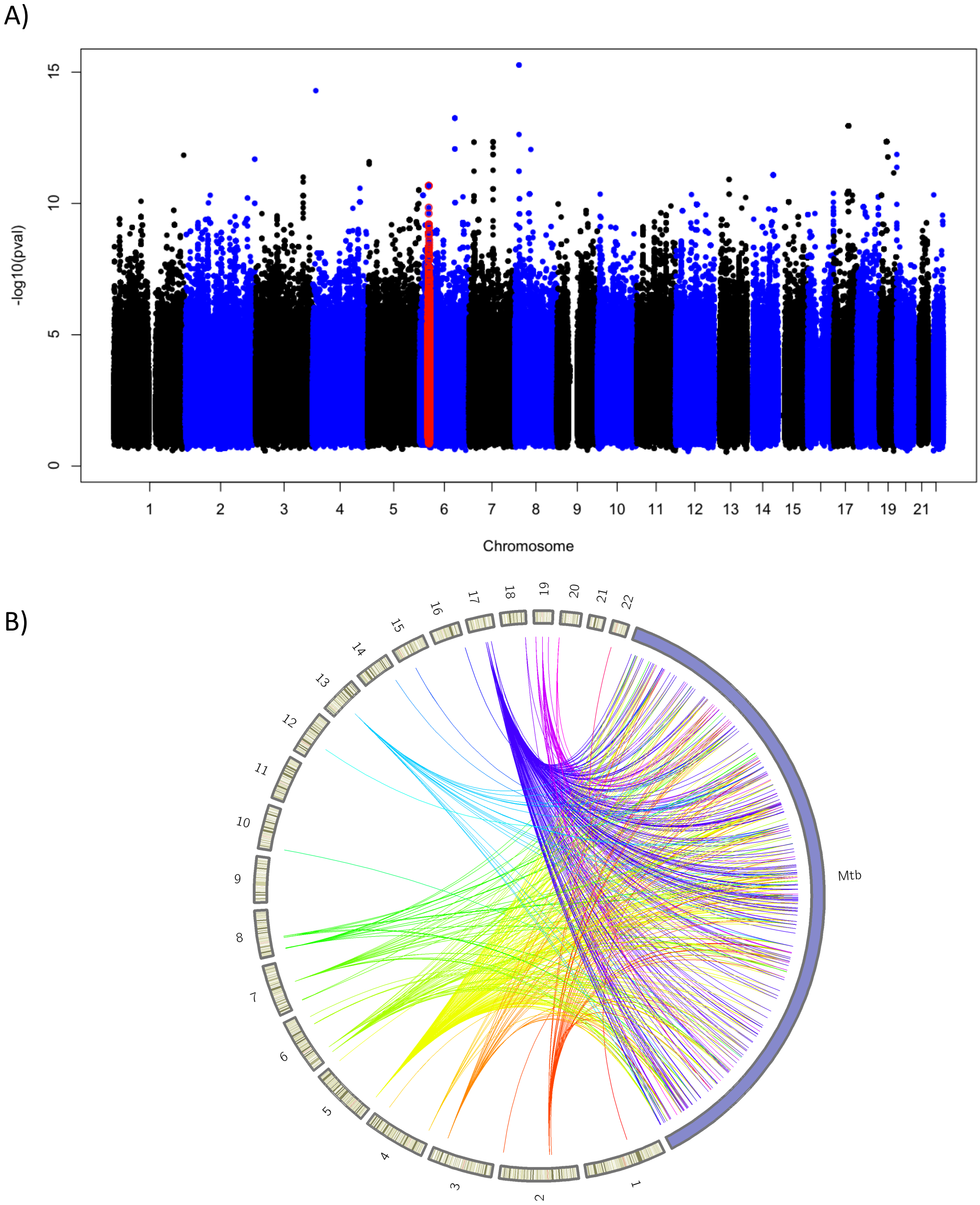


Figure 2

Results from the genome-to-genome comparison of host and pathogen data.

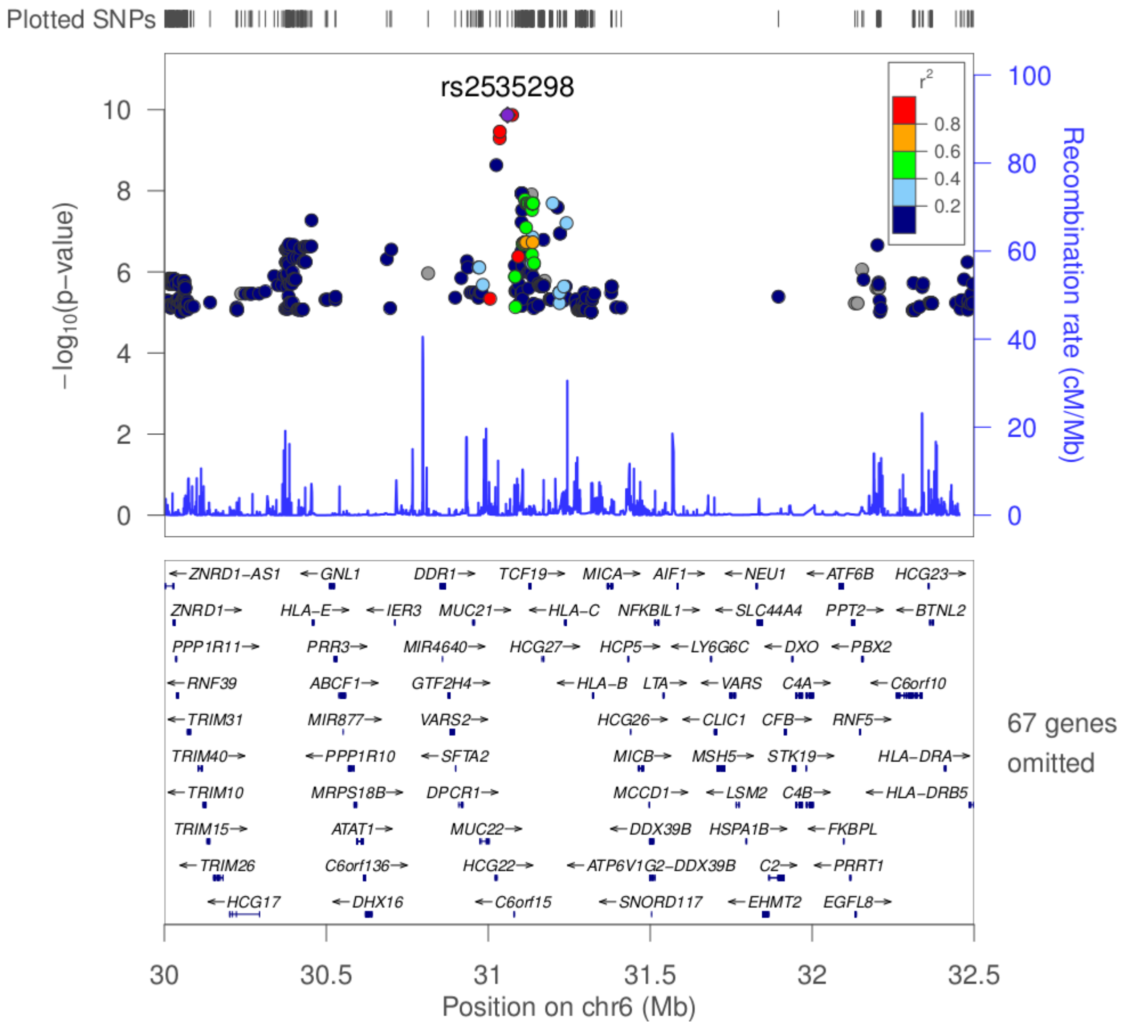
(A) A Manhattan plot showing the *–log10(P-value)* for each human variant. Results are plotted by chromosomes with alternating black and blue colouring. The MHC locus is highlighted in red. (B) A circus plot showing the associations between host and pathogen genomes. The Pathogen genome is coloured in blue and the human chromosomes in cream (not to scale). A link is drawn between the two genomes where a variant at the corresponding positions passed the 10^{-10} association cut-off value.



Supplementary figure 1

Host genetic regions identified as having interactions with the *M. tuberculosis* genome

(a) Human leukocyte antigen (HLA) region



Supplementary table 1
The distribution of *M. tuberculosis* lineages within each Human PCA-based cluster group

M. tuberculosis Lineage	Human group 1	Human Clus 2	Human Cluster 3	Total
1	232 (45.7%)	16 (12.0%)	6 (7.6%)	254 (35.3%)
2	215 (42.3%)	76 (57.1%)	50 (63.3%)	341 (47.4%)
3	5 (1.0%)	2 (1.5%)	1 (1.3%)	8 (1.1%)
4	56 (11.0%)	39 (29.3%)	22 (27.8%)	117 (16.3%)
Total	508	133	79	720

Supplementary table 2
Frequency and significant ($p<10^{-6}$) association of Human leukocyte antigen (HLA) types to *M. tuberculosis* genotypes

HLA Type	P value	Frequency
HLA_DQA1_06	1.26E-08	0.194
HLA_DQA1_0601	1.26E-08	0.194
HLA_B_15	1.49E-07	0.218
HLA_DPA1_02	2.78E-07	0.593
HLA_DQB1_0301	9.69E-07	0.233
HLA_DPA1_01	2.09E-06	0.332
HLA_DRB1_1202	4.40E-06	0.220
HLA_DRB1_12	6.07E-06	0.224
HLA_A_11	8.13E-06	0.447

Chapter 9

Discussion and Conclusion

9 Discussion and Conclusion

9.1 Discussion

This thesis focuses, for the most part, on the analysis of Next Generation Sequencing (NGS) data in the context of TB. **Chapter 2** worked to establish the fidelity of the data from two NGS technologies, Illumina MiSeq and Ion PGM platforms. I have shown the variability between technical and biological replicates to be negligible. Phylogenetic reconstruction revealed interleaved clustering of biological and technical replicates. This high similarity indicates the lack of systematic bias in sequencing results caused by the extraction process. A differential GC%-dependant coverage bias was observed between the Illumina and Ion PGM platforms. While both platforms displayed drops in coverage across high-GC regions, its effect was more drastic with the PGM. This lower level affected the calling of large indels, with more false positives using the PGM, however most known *Mtb* drug resistant genes were well characterised using both technologies. The rapid development of NGS technologies has led to a significant decrease in the cost and of throughput sequencing. As demonstrated here, we are now at a stage where sequencing is no longer the bottleneck. Removing or decreasing the time required at the culturing step should be a main focus of future research. Platforms such as minION may in the future enable sequencing to be performed on site at clinics. A rigorous study of the error rates and profiling capabilities such as described above should also be performed on other new technologies as they become available.

The study also highlighted the difference between two popular resistance profiling software, Mykrobe TB Profiler and TB predictor. The differences in the underlying drug resistance mutation database caused a number of false positives using the Mykrobe TB Profiler. However, international efforts, such as the ReSeqTB project, are aiming to consolidate a comprehensive single library of all variants, which will eventually result in all such software converging. However, none of the tools evaluated use large deletions or insertions for profiling, but this would be a straightforward implementation of looking for decreases in coverage. This is required because variants such as a deletion on the *thyA-dfrA* genes causing PAS resistance was observed, but would be undetected by conventional approaches.

There is a need for a better understanding of mutations involved in *Mtb* drug resistance, including for any new drugs. To this effect, in **Chapter 3** I applied GWAS methods to identify such resistance mutations and loci. To develop the methodological approach, I used whole genome sequence data from 127 clinical isolates with corresponding DST and MIC data for the first line drugs rifampicin, isoniazid, ethambutol and streptomycin. Using GWAS and phylogenetic approaches I found mutations in known resistance genes, thus validating their use for drug resistance variant discovery. Additionally, the effects of the variants on protein structure stability and drug binding were modelled in the RpoB and KatG proteins. A high correlation was found between the level of resistance, measured by MIC, and the distance to the drug binding site. This information could be used to predict the effect of novel variants on drug resistance. For example, in theory it may be possible to perform a genome-wide assessment of the effects of all possible mutations on protein structures. Drug resistance can be conferred by changes to protein stability, protein-protein interactions and ligand binding. There are many programs which can collect metrics on the changes of these properties caused by mutation. A potential route of future investigations could be the application of deep neural networks ,which could model the effect of multiple parameters, to these metrics to predict drug resistance variants. However, many of the protein structures for drug targets are currently unavailable or do not characterise the whole protein.

After successful implementation of the GWAS method (**Chapter 3**), in **Chapter 4** I applied the analysis pipeline on a global set of 6,465 strains. This dataset combined publically available data and our own in-house sequencing, and represents one of the largest studies of drug resistance in *Mtb* to date. The large sample size (n>6,000) meant I was able to detect resistance variants at low minor allele frequency, as well as the genes involved by combining the rare alleles (as applied in **Chapter 3**). Large sample sizes are needed to perform per variant approaches, and thereby provide a higher resolution to identify causal variants. This may be difficult to achieve for new forms of resistance where samples sizes are small, but the effects sizes are likely to be large. In my work, lineages 1, 2, 3 and 4 were represented, allowing for the detection of strain-specific effects along with increasing the chance to pick up mutations appearing convergently across all lineages using the *PhyC* approach. To look for novel variants the results from the GWAS and *PhyC* methods were compared to the well-established resistance variant list from the T*BProfiler* database. A large number of mutations were found in well-established resistance loci such as *rpoB* and *katG.* Additionally, I report several novel mutations which were not present in the database. These will need to be validated using allelic exchange experiments, and is beyond the scope of my thesis. However, I looked at the effect of adding the new mutations on the sensitivity of the database and found sensitivity gains ranging from 1% (isoniazid/rifampicin) to 55% (PAS) para-amino salicylic acid. Many small indels and large deletions were found in drug activating enzymes. Using small indels increased the sensitivity from 20% to 40% for PAS, adding large deletions increased this 65%. By including these new mutations, we can improve predictive accuracy of mutation libraries and bring *in-silico* prediction a step closer to application in a clinical setting. By developing large databases across clinical settings of well characterised *Mtb* with whole genome sequences and DSTs, it may be possible apply machine learning methods to detect mutations and through a learning process update mutation libraries[1]. Analyses involving large sample sizes are likely to be robust to errors in DSTs, and the use of MIC values may assist issues with resistance cut-offs.

To date, mutation libraries have focused on SNPs and small indels in drug targets or activators. Although relatively rare for some pro-drugs such as INH, large indels seem to play a major role in PAS and ETH resistance and calls for the integration large deletion calling into existing profiling tools. Association was found between the XDR phenotype and the efflux pump *drrA* which has been reported to cause resistance to anitibiotics[2]. Another efflux pump (*Rv2688c*) was associated with fluoroquinolone resistance[3]. This gene has been reported to cause resistance to ciprofloxacin and moxifloxacin when expressed in *M. smegmatis*[3]. This highlights the importance of the efflux pumps in conferring antibiotic resistance and could help to explain the resistance in isolates which do not have mutations in the drug targets or activators.

In **Chapter 3** a correlation was found between the number of mutations in drug resistance genes with the MIC for that drug. Although I did not have the MIC values for the 6,545 used in Chapter 4, I looked for a correlation between the distribution of odds ratio in a gene and information on the known levels of resistance conferred by that gene. Unsurprisingly, I found a positive correlation between the median odds ratio and the level of resistance conferred by the gene. Whilst this data is very preliminary, the odds ratios of resistance mutations appears to be an epidemiological surrogate of levels of resistance.

In summary, the work in **Chapter 4** has identified potentially new variants, which would require validation work in laboratory experiments. Whilst this is a slow and expensive process using *Mtb*, a number of surrogate models have been proposed[4,5]. *M. aurum* has been proposed as a good surrogate model for *Mtb* due to the similarities in cell wall lipid content and drug sensitivity profile. A good characterisation of the genome is required for allelic exchange experiments; however, no draft reference sequence has been published for *M. aurum*. In **Chapter 5**, to facilitate its use as a model organism, I have analysed sequence data of *M. aurum* (NCTC 10437) and assembled a draft reference genome. Comparison of the *Mtb* reference sequence (H37Rv) with our assembly revealed a high degree of similarity between drug resistance genes of *Mtb* and their homologues in *M. aurum*. The genome of *M. aurum* is significantly larger than *Mtb* and, as such, was found to contain 2,090 genes which were not found in *Mtb*. These genes could potentially influence allelic exchange experiments and high throughput drug screening by metabolising potential compounds or by providing alternate resistance mechanisms and should be considered in experimental design. The published draft reference should aid the development of *M. aurum* as a surrogate model for *Mtb* and aid in the development of new drugs and elucidation of resistance mechanisms. An interesting finding was the presence of copy number variants of *katG* and *embB*, which needs to be considered when analysing derivatives of isoniazid and ethambutol.

Whilst drug resistance can be explained by mutations a select few genes or pathway, phenotypes such as virulence are dictated by a multitude of proteins and pathways, which interact directly or indirectly interact with the host. Although we have a good understanding of the genetic variation in *Mtb*, ~10% of the genome corresponding to the *pe* and *ppe* genes is routinely ignored. Standard mapping techniques do not perform well with the repetitive sequences present in these genes and as a result are discarded. In **Chapter 6**, to improve characterisation of the *pe/ppe* gene families I performed genome assembly on a set of 518 isolates with high depth of coverage. All isolates had >70% fully assembled *pe/ppe* genes, and the remaining genes were > 90% assembled. By comparing to the H37Rv reference strain 5,853 SNPs were found in the *pe/ppe* genes, equating to roughly 11.6% of the total number of SNPs. Phylogenetic analyses pointed to a region surrounding the *pe_pgrs3* and *pe_pgrs4* genes causing anomalous clustering of strains. A large number of variants were observed in this region across the dataset. These genes share a highly similar sequence and could potentially recombine. I hypothesised that recombination between these genes contributes towards the large amount of sequence diversity seen in this region. In **Chapter 7** I used PacBio sequencing to confirm the presence of a large structural variant occurring in a lineage 2 strain. This variant caused the insertion of a large amount of sequence in between the *PE_PGRS3* and *PE_PGRS4* genes. Sequence annotation reported two small open reading frames within this new region. This work demonstrates that, in general, the *pe/ppe* genes show a lineage specific pattern of variation, and intragenic recombination may contribute towards this variation. This variation may contribute towards phenotypes which differ between strains, such as transmissibility and virulence, especially since the *pe/ppe* genes are thought to be in contact with host cells. The *pe/ppe* genes have been proposed as vaccine candidates. This work has highlighted the presence of selection acting on a subset of these genes and needs to be considered in the development process.

The work so far has focused on the analysis of DNA sequence, methylation of DNA can modulate the effects of genomic variation. In **Chapter 7** I sought to use PacBio SMRT sequencing to characterise methylation in *Mtb* and to confirm observations described in **Chapter 6.** The dataset used for PacBio sequencing included strains from diverse genetics to capture as much genetic variation as possible. Sixteen isolates were sequenced and supplemented with two sequences from the ENA database. These strains were selected to represent the diversity seen in *Mtb*, with all lineages except lineage 7 represented. Future work should consider lineage 7, as it is an intermediate strain. Three candidate motifs were identified, although not all motifs were methylated in each sample. The differential methylation of motifs in different strains prompted me to look for mutations which might lead to the lack of methylation and analyse these in terms of their distribution across the lineages. Six potential loss of function (LOF) mutations were identified and their distributions were characterised in the >6000 strains described in **Chapter 4**. Surprisingly, loss of function in methyltransferase genes is common. All lineage 3 isolates and 71% of lineage 4 isolates contained a LOF mutation in the MTase associated with the GATN4RTAC motif. All lineage 2 isolates harboured a LOF mutation in the MTase (*MamA*) associated with the CTCCAG motif. Previous reports have shown that strains lacking a functional copy of *MamA* had a decreased survival rate in hypoxic conditions[6]. Interestingly, “modern” Beijing lineages have been reported to upregulate the DosR regulon[7] which is regulated in response to hypoxia[8]. Whether the co-occurrence of the LOF in MTase and upregulation of the DosR regulon is simply chance or whether they are linked would be an interesting question to investigate. The effects of a MamA LOF in H37Rv strains have been investigated, where this strain-type have normal expression of the DosR regulon[6]. Application of a Beijing strain might lead a different effect on the survival rate. The thesis work puts methylation in *Mtb* in a global context and demonstrates the strain and lineage specific methylation patterns. Little functional work has been performed to elucidate the effect of methylation in TB, and this work highlights the importance role that methylation could play in explaining phenotypic differences between strains. Future work could include investigating gene expression and the heritability of methylation. Strains could be whole genome sequenced using PacBio technology and whole transcriptome sequenced using an Illumina platform. This would allow investigation of the direct effect of methylation on transcription. Additionally, PacBio sequencing of transmission clusters could provide insights into the heritability of methylation in *Mtb*.

In **Chapter 8**, I assessed the potential of using host and pathogen genomic data to identify host-pathogen interactions. I hypothesised that the lack of replication in GWASs could be due to the differential strains circulating globally. Previous studies have shown a close co-evolution between humans and *Mtb*[9]. This evolutionary effect could lead to adaptation of *Mtb* or resistance of humans to historically circulating strains. I performed an analysis of host genotyping and pathogen WGS data from Thailand to identify potential interactions, as measured by the co-occurrence of mutations on both genomes in a genome-to-genome analysis. Using a GWAS, using the pathogen genotypes as a phenotype I identified highly significant associations between subclades of *Mtb* and human variants. I also find, to a lesser extent, association between the HLA region and lineage 1. These putative interactions could potentially represent selection acting on the pathogen to evade the host immune system, however validation is required. Using a similar dataset from another population with similar pathogen strains in circulation (e.g. Vietnam) could provide the support required before functional work is performed. Work is underway to perform a similar analysis within a Vietnamese TB cohort, and could lead to validation of findings.

9.2 Conclusions

This thesis presents the analysis of *Mtb* and host genomic data to characterise the variation and its downstream effects. **Chapters 2** to **5** focus on the evolution of drug resistance and the applicability of NGS to predict drug resistance. **Chapters 6** and **7** focus on the less well characterised variation of the *pe/ppe* genes and methylome which could correlate to phenotypic differences between strains. Finally, **Chapter 8** integrates host data to look for specific interactions between the two genomes. I hope that this data will enable researchers to answer questions concerning the diversity of *Mtb* and that results from this work will contribute towards our understanding of this complex pathogen.

9.3 The future of Tb genomic analysis

The cost of sequencing has fallen at a faster rate than Moore's law predictions[10] and is likely to continue along this trajectory in the next couple of years. This price drop has enabled sequencing of thousands of isolates as demonstrated in this thesis. The bottleneck to scaling up *Mtb* sequencing projects currently lies at the culture step. Whilst significant strides have been made towards the achievement of this goal[11,12], a cost effective and high throughput solution is still lacking and further development is needed. We have shown the value of having big datasets to discover novel drug resistance variants. As national TB programs start to adopt WGS as a diagnostic tool a large amount of data will be collected. This not only has an impact at the patient level, but serves to provide an epidemiological view of TB at a national and international level. Work should focus on building a platform for real-time analyses of new sequence data to provide useful information, such as drug resistance, to clinicians. Additionally, sequence data generated in clinics could be sent to a centralised database along with meta data such as location and date of collection. Here, new data could be integrated with all other sequences in real-time to update epidemiological metrics which could help with developing control measures. Results from DSTs could be fed in and a real-time GWAS could be envisioned whereby, p-values are updated on-the-fly to create a live drug resistance database. As more data is generated, the sensitivity and specificity of the database will increase too, thereby encouraging more countries/regions to take part and generating a positive feedback loop. Geographical data could also provide insights into routes of transmission and help TB programs identify high risk areas and where to focus efforts. As host-based therapies advance, so will the need to characterise host-pathogen interactions[13]. Data from patients could optionally be collected at the same time as the pathogen and used in a similar analysis as described in **Chapter 8**. This could potentially delineate the molecular interaction mechanisms and lead to a more effective host-directed therapy. Future advances in technology and software will bring prospect of sequencing in the clinic closer to realisation, helping with patent management and decreasing the burden on public health services. This in turn will facilitate the epidemiological surveillance of TB on a national and international level and is likely to lead to greater insights into the disease and control measures, leading to the fulfilling of WHO targets for eradication.

References

1. Yang, Y., Niehaus, K. E., Walker, T. M., Iqbal, Z., Walker, A. S., Wilson, D. J., Peto, T. E., Crook, D. W., Smith, E. G., Zhu, T. & Clifton, D. A. Machine Learning for Classifying Tuberculosis Drug-Resistance from DNA Sequencing Data. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btx801
2. Choudhuri, B. S., Bhakta, S., Barik, R., Basu, J., Kundu, M. & Chakrabarti, P. Overexpression and functional characterization of an ABC (ATP-binding cassette) transporter encoded by the genes drrA and drrB of Mycobacterium tuberculosis. *Biochem. J.* **367,** 279–85 (2002).
3. Pasca, M. R., Guglierame, P., Arcesi, F., Bellinzoni, M., De Rossi, E. & Riccardi, G. Rv2686c-Rv2687c-Rv2688c, an ABC Fluoroquinolone Efflux Pump in Mycobacterium tuberculosis. *Antimicrob. Agents Chemother.* **48,** 3175–3178 (2004).
4. Gupta, A. & Bhakta, S. An integrated surrogate model for screening of drugs against Mycobacterium tuberculosis. *J. Antimicrob. Chemother.* **67,** 1380–91 (2012).
5. Agrawal, P., Miryala, S., Varshney, U., Mallick, B., Yamunadevi, S. & Sang, P. Use of Mycobacterium smegmatis Deficient in ADP-Ribosyltransferase as Surrogate for Mycobacterium tuberculosis in Drug Testing and Mutation Analysis. *PLoS One* **10,** e0122076 (2015).
6. Shell, S. S., Prestwich, E. G., Baek, S.-H., Shah, R. R., Sassetti, C. M., Dedon, P. C. & Fortune, S. M. DNA Methylation Impacts Gene Expression and Ensures Hypoxic Survival of Mycobacterium tuberculosis. *PLoS Pathog.* **9,** e1003419 (2013).
7. Reed, M. B., Gagneux, S., Deriemer, K., Small, P. M. & Barry, C. E. The W-Beijing lineage of Mycobacterium tuberculosis overproduces triglycerides and has the DosR dormancy regulon constitutively upregulated. *J. Bacteriol.* **189,** 2583–9 (2007).
8. Sherman, D. R., Voskuil, M., Schnappinger, D., Liao, R., Harrell, M. I. & Schoolnik, G. K. Regulation of the Mycobacterium tuberculosis hypoxic response gene encoding alpha -crystallin. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 7534–9 (2001).
9. Gagneux, S. Host-pathogen coevolution in human tuberculosis. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367,** 850–9 (2012).
10. Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D. J., Salichos, L., Zhang, J., Weinstock, G. M., Isaacs, F., Rozowsky, J. & Gerstein, M. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.* **17,** 53 (2016).
11. Brown, A. C., Bryant, J. M., Einer-Jensen, K., Holdstock, J., Houniet, D. T., Chan, J. Z. M., Depledge, D. P., Nikolayevskyy, V., Broda, A., Stone, M. J., Christiansen, M. T., Williams, R., McAndrew, M. B., Tutill, H., Brown, J., Melzer, M., Rosmarin, C., … Breuer, J. Rapid Whole Genome Sequencing of M. tuberculosis directly from clinical samples. *J. Clin. Microbiol.* JCM.00486-15- (2015). doi:10.1128/JCM.00486-15
12. Votintseva, A. A., Bradley, P., Pankhurst, L., del Ojo Elias, C., Loose, M., Nilgiriwala, K., Chatterjee, A., Smith, E. G., Sanderson, N., Walker, T. M., Morgan, M. R., Wyllie, D. H., Walker, A. S., Peto, T. E. A., Crook, D. W. & Iqbal, Z.

Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. *J. Clin. Microbiol.* **55,** 1285–1298 (2017).

13. Wallis, R. S. & Hafner, R. Advancing host-directed therapy for tuberculosis. *Nat. Rev. Immunol.* **15,** 255–263 (2015).