

LONDON
SCHOOL of
HYGIENE
& TROPICAL
MEDICINE



Bioinformatic analysis of *Mycobacterium tuberculosis*
whole genome data

Francesc Coll I Cerezo

2015

Thesis submitted in accordance with the requirements
for the degree of Doctor of Philosophy of the
University of London

Pathogen Molecular Biology Department
Faculty of Infectious and Tropical Diseases
London School of Hygiene & Tropical Medicine

Funded by Bloomsbury Colleges PhD Studentships

Declaration

I, Francesc Coll I Cerezo, 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 the thesis.

Signed

A black rectangular box redacting the signature of the author.

Date 10/12/2014

ABSTRACT

Tuberculosis (TB) caused by bacteria of the *Mycobacterium tuberculosis* complex (MTBC) is the second major cause of death from an infectious disease worldwide. Recent advances in DNA sequencing are leading to the ability to generate whole genome information of clinical isolates of MTBC. The objectives of this work include developing bioinformatic tools for processing and making accessible MTBC genomic data, as well as the identification of informative genetic markers, both strain-specific and associated with drug resistance (DR), to barcode MTBC isolates in research and clinical settings.

SpolPred software was developed to accurately predict the spoligotype from raw sequence reads, and used to bridge the gap between classical genotyping and high-throughput sequencing. A genome variation discovery pipeline was implemented to derive genomic polymorphisms from MTBC raw sequence data. This pipeline was applied to >1,500 publicly available isolates and the characterised genomic variation hosted in *PolyTB*, a web-based tool where genetic variants can be investigated using a genome browser, a world map showing their global allele distribution, and an additional phylogenetic view. An extensive repertoire of strain-specific mutations was identified, of which a subset was proposed to accurately discriminate known MTBC circulating strains. A curated list of DR associated mutations was compiled from the literature and their diagnostic accuracy for predicting phenotypic resistance assessed. In addition, potentially novel genes involved in DR were discovered by applying genome-wide association approaches to a global population of more than 2,500 MTBC strains.

Whole genome sequencing (WGS) promises to be transformative for the practice of clinical microbiology, and the rapidly falling cost and turnaround time mean that this will become a viable technology in clinical settings. In this new paradigm, the presented work will facilitate the transition to and applications of WGS in clinical settings as an important tool for TB control.

ACKNOWLEDGEMENTS

I would like to thank all the people who helped me in one way or another during the course of my PhD. I am very grateful to the Bloomsbury Colleges PhD Studentships for funding my PhD. This thesis would not have been possible without the support and enthusiasm of my supervisors, Dr Taane G Clark in the London School of Hygiene and Tropical Medicine, and Nigel Martin in Birkbeck College, for which I will be extremely grateful. I would like to acknowledge the support of Dr Ruth Mcnerney within the London School of Hygiene and Tropical Medicine, especially for her academic input and constructive criticism. In the same terms, I acknowledge the contribution of all our collaborators.

I am deeply grateful with the colleagues I have worked with within Taane's group during these three years; for their continuous advice, contribution to my work and good times outside the office. Last but not least, I would like to thank my family and friends in Valencia, for their hearty welcomes every time I went back home; and the friends I met in London with whom I share the best memories in London.

TABLE OF CONTENTS

ABSTRACT	3
ACKNOWLEDGEMENTS.....	5
LIST OF FIGURES.....	8
LIST OF TABLES.....	9
LIST OF SUPPLEMENTARY TABLES	9
LIST OF SUPPLEMENTARY FIGURES.....	10
LIST OF ABBREVIATIONS	11
1 INTRODUCTION.....	16
1.1 Tuberculosis disease and etiological agent.....	16
1.2 Impact of WGS on TB research and clinical applications	22
1.3 Next-generation sequencing technologies	27
1.4 Whole genome sequence analyses.....	28
1.5 Research aims and objectives	32
1.6 Description of the thesis and contributions.....	33
1.7 Description of the data sets	38
2 IDENTIFYING AND VISUALISING GENOMIC VARIATION	41
2.1 INTRODUCTION.....	41
2.2 METHODS.....	43
2.2.1 Genomic variation discovery pipeline from Illumina paired-end sequence data.....	43
2.2.2 Genomic variation discovery pipeline from complete genomes	46
2.2.3 Population structure	46
2.2.4 <i>PolyTB</i> software architecture.....	46
2.3 RESULTS	51
2.3.1 Polymorphisms detected and incorporated into <i>PolyTB</i>	54
2.3.2 <i>PolyTB</i> and its applications	57
2.4 DISCUSSION	62
3 STRAIN TYPING USING WHOLE GENOME SEQUENCES	67
3.1 INTRODUCTION.....	67
3.2 METHODS.....	70
3.2.1 Whole genome datasets and sequence analysis	70
3.2.2 <i>In silico</i> determination of spoligotypes patterns from short genomic sequences	71
3.2.3 <i>In silico</i> determination of lineages by regions of difference	73
3.2.4 Phylogenetic analysis	75
3.2.5 Identification of clade-specific SNPs and selection of the minimal informative set.....	75
3.3 RESULTS	76
3.3.1 <i>In silico</i> prediction accuracy of spoligotype patterns.....	76
3.3.2 Population structure of the global collection of MTBC strains	78
3.3.3 Identification of lineage and sub-lineage specific SNPs and selection of the minimal informative set.....	84
3.3.4 Validation of the proposed SNP-typing system and comparison to other SNP sets.....	86

3.4 DISCUSSION	87
4 A WHOLE-GENOME SEQUENCING APPROACH FOR DRUG RESISTANCE PROFILING ..	93
4.1 INTRODUCTION.....	93
4.1.1 Drug resistance: a threat to disease control.....	93
4.1.2 Review of mechanisms of drug resistance in <i>Mycobacterium tuberculosis</i> ..	94
4.1.3 Available diagnostic tests for drug resistant tuberculosis	98
4.1.4 Whole-genome sequencing for the detection of drug resistance	99
4.2 METHODS.....	100
4.2.1 Mutation library.....	100
4.2.2 Sequence data and drug susceptibility testing	101
4.2.3 Rapid mutation detection and the <i>TB Profiler</i> Online tool	103
4.2.4 Comparison with existing tools.....	105
4.3 RESULTS	106
4.3.1 Validation of mutation library.....	106
4.3.2 Comparison with commercial tests	110
4.3.3 Comparison with other drug resistance databases	112
4.3.4 Online tool for predicting drug resistance and lineage information from sequenced isolates.....	115
4.4 DISCUSSION	116
5 IDENTIFICATION OF NOVEL DRUG RESISTANCE ASSOCIATED LOCI USING GENOME- WIDE ASSOCIATION ANALYSIS.....	122
5.1 INTRODUCTION.....	122
5.2 METHODS.....	125
5.2.1 Dataset, raw sequence alignment and SNP calling	125
5.2.2 Phylogenetic reconstruction and population structure	127
5.2.3 Phylogenetic convergence test for selection	127
5.2.4 Genome-wide association analysis	128
5.3 RESULTS	129
5.3.1 Population structure	129
5.3.2 Phenotypic drug resistance explained by known candidate genes	132
6 DISCUSSION AND FURTHER WORK	150
REFERENCES.....	162
SUPPLEMENTARY MATERIAL	187

LIST OF FIGURES

Figure 1.1 Estimated TB incidence rates in 2012	16
Figure 1.2 Percentage of new TB cases with MDR-TB in 2012.....	17
Figure 1.3 TB pathogenesis.....	19
Figure 1.4 MTBC evolutionary history	24
Figure 1.5 Fastq format.....	29
Figure 1.6 Alignment visualisation with a SNP and sequencing errors	31
Figure 2.1 Genome variation discovery pipeline	44
Figure 2.2 Genome Browser View web architecture	48
Figure 2.3 Map View web architecture.....	49
Figure 2.4 Phylogenetic View web architecture	51
Figure 2.5 <i>R</i> AxML phylogenetic tree built for all 1,470 MTBC isolates (colour-coded by spoligotype)	52
Figure 2.6 <i>R</i> AxML phylogenetic tree built for all 1,470 MTBC isolates (colour-coded by geographical location)	53
Figure 2.7 SNP frequency bar plot	54
Figure 2.8 Box-and-Whisker SNP density plots by gene functional categories.....	55
Figure 2.9 Indel frequency bar plot	56
Figure 2.10 Box-and-Whisker indel density plots by gene functional categories	56
Figure 2.11 Polymorphisms at the <i>rpoB-rpoC</i> region (Browser View)	58
Figure 2.12 SNP associated with lineage 1 (EAI) in Tanzanian and Karonga-Malawian populations (Map view)	60
Figure 2.13 SNP-based neighbour-joining phylogenetic tree of 140 isolates belonging to four different locations (Phylogenetic view)	61
Figure 3.1 Experimental spoligotyping technique	72
Figure 3.2 <i>In silico</i> spoligotyping	73
Figure 3.3 Dendrogram for 51 Ugandan isolates constructed using 7k SNPs	77
Figure 3.4 Global phylogeny of 1,601 MTBC isolates.....	79
Figure 3.5 Distribution of lineage-specific SNPs across gene functional categories	85
Figure 4.1 The <i>TB profiler</i> tool	104
Figure 4.2 Mutations associated with MDR-TB found in phenotypically MDR strains	108
Figure 4.3 Mutations associated with XDR-TB found in phenotypically XDR strains...	109
Figure 4.4 Inferred analytical accuracies of the DR curated mutation library and three commercial molecular tests for resistance	111
Figure 4.5 Diagnostic performance of the curated library versus alternative drug resistance mutation databases.....	113
Figure 5.1 Phylogeny of the global drug resistance MTBC samples	130
Figure 5.2 Phylogeny of the global drug resistance MTBC samples colour-coded by drug resistance status	131
Figure 5.3 Percentage of resistance cases explained by known DR markers.....	133
Figure 5.4 Principal Components Analysis for the global drug resistance data set	135
Figure 5.5 Locus GWAS results for isoniazid	137
Figure 5.6 Locus GWAS results for rifampicin.....	138
Figure 5.7 Locus GWAS results for ethambutol	138
Figure 5.8 Locus GWAS results for pyrazinamide	138

Figure 5.9 Locus GWAS results for streptomycin.....	139
Figure 5.10 Locus GWAS results for ofloxacin	139
Figure 5.11 Locus GWAS results for ethinamide.....	139
Figure 5.12 Locus GWAS results for amikacin.....	140
Figure 5.13 Locus GWAS results for capreomycin	140
Figure 5.14 Locus GWAS results for kanamycin.....	140
Figure 5.15 <i>PhyC</i> results.....	143

LIST OF TABLES

Table 1.1 Commercially available second generation sequencing platforms.....	27
Table 1.2 Research papers included in this thesis in chronological order	34
Table 1.3 Summary of all WGS TB studies	38
Table 1.4 Composition of the four WGS data sets.....	39
Table 2.1 Comparison table of TB WGS genomic databases	64
Table 3.1 Phylogenetically informative deletions.....	74
Table 3.2 Lineage characteristics	80
Table 3.3 MTBC lineages and sub-lineages.....	81
Table 3.4 SNP typing systems comparison	87
Table 4.1 Summary of mutations included in the curated drug resistance mutation library.....	100
Table 4.2 Accuracy of whole genome drug resistance analysis compared to reported resistance phenotype	107
Table 4.3 Potentially novel DR mutations identified by <i>TB profiler</i>	115
Table 5.1 Summary of the global drug resistance dataset.....	126
Table 5.2 Previously unreported mutations in candidate genes	134

LIST OF SUPPLEMENTARY TABLES

Supplementary Table 1 <i>Mtb</i> complete genomes used to generate a set of validated SNP, indel and large deletion loci.	187
Supplementary Table 2 Experimental spoligotyping and <i>SpolPred</i> results for the 51 Ugandan samples.....	188
Supplementary Table 3 Lineage composition of the WGS data set 2 populations.....	189
Supplementary Table 4 The set of 62 phylogenetically informative SNPs for MTBC typing	190
Supplementary Table 5 Lineage predictions for a set of reference genomes	193
Supplementary Table 6 Lineage, specific SNPs at drug resistance genes	194
Supplementary Table 7 Non-synonymous lineage-specific SNPs at known epitopes in H37Rv.....	197
Supplementary Table 8 Sub-lineage proportions observed in the Russian dataset (n=850) (Casali <i>et al.</i> 2014) when using the 62-SNPs classification scheme.....	199
Supplementary Table 9 Probable cases of Karonga-Malawian mixed samples.....	200

Supplementary Table 10 Locus-based GWAS top hits	201
Supplementary Table 11 Operon-based GWAS top hits	212

LIST OF SUPPLEMENTARY FIGURES

Supplementary Figure 1 Proportion of missed call across all samples	217
Supplementary Figure 2 SNP Allele Frequency Spectrum	219
Supplementary Figure 3 Global phylogeny of 1,601 MTBC isolates colour, coded by spoligotype	220
Supplementary Figure 4 Global phylogeny constructed using the 62 SNP typing system	221
Supplementary Figure 5 Global phylogeny constructed using the 45 SNP typing system proposed by Filliol <i>et al.</i> 2006	222
Supplementary Figure 6 Global phylogeny constructed using the 93 SNP typing system proposed by Comas <i>et al.</i> 2009	223
Supplementary Figure 7 Global phylogeny constructed using the 71 SNP typing system proposed by Homolka <i>et al.</i> 2012	224
Supplementary Figure 8 Loci involved in drug resistance	225
Supplementary Figure 9 Cumulative sensitivity and specificity of drug resistance markers	235
Supplementary Figure 10 Diagnostic accuracy across populations	246

LIST OF ABBREVIATIONS

AA	Amino Acid
AMI	Aminoglycosides
AMK	Amikacin
BAM	Binary sequence Alignment/Map format
BCG	Bacille Calmette-Guérin
BGI	Beijing Genomics Institute
bp	base pairs
CAP	Capreomycin
CAS	Central Asian
Chg.	Change
ChIP-seq	Chromatin Immunoprecipitation followed by Sequencing
CIP	Ciprofloxacin
CJC	Cross-Junction Contig
CNV	Copy Number Variant
Coor.	Coordinate
CYS	Cycloserine
dbSNP	Single Nucleotide Polymorphism Database
DNA	Deoxyribonucleic Acid
DOC	Depth Of Coverage
DR	Drug Resistance
DST	Drug Susceptibility Testing
EAI	East African-Indian
EMB	Ethambutol
ENA	European Nucleotide Archive
ETH	Ethionamide
FLQ	Fluoroquinolones
fLS-SNPs	filtered Lineage Specific Single Nucleotide Polymorphisms
GMTV	Genome-based Mycobacterium Tuberculosis Variation
GWAS	Genome-Wide Association Study

HGT	Horizontal Gene Transfer
HIV	Human Immunodeficiency Virus
indel	insertion and deletion
INH	Isoniazid
KAN	Kanamycin
KAUST	King Abdullah University of Science and Technology
kb	kilobases
LAM	Latin-American
LEVO	Levofloxacin
LiPA	Line Probe Assays
LSP	Large Sequence Polymorphisms
LS-SNPs	Lineage Specific Single Nucleotide Polymorphisms
Mb	Mega/Million base pairs
MDR	Multidrug Resistance
MDR-TB	Multidrug Resistance Tuberculosis
MGDD	Mycobacterial Genome Divergence Database Mycobacterial Interspersed Repetitive Units - Variable Number
MIRU-VNTR	Tandem Repeat
MLPA	Multiplex Ligation-dependent Probe Amplification
MLST	Multi-locus sequence typing
MOL-PCR	Multiplexed Oligonucleotides Ligation PCR
MOX	Moxifloxacin
Mtb	Mycobacterium tuberculosis
MTBC	Mycobacterium tuberculosis complex
NCBI	National Center for Biotechnology Information
ND	Not determined
NGS	Next Generation Sequencing
NonSyn	Non-Synonymous
NS	Non-Synonymous
nsSNP	Non-Synonymous Single Nucleotide Polymorphism
NT	Nucleotide

Num.	Number
OFX	Ofloxacin
OLC	Overlap/Layout/Consensus
PAS	Para-aminosalicylic acid
PATRIC	Pathosystems Resource Integration Center
PCA	Principal Components Analysis
PCR	Polymerase Cycling Assembly
PCs	Principal Components
PZA	Pyrazinamide
QRDR	Quinolone Resistance-Determining Region
RD	Regions of Difference
RFB	Rifabutin
RFLP	Restriction Fragment Length Polymorphism
RMP	Rifampicin
RNA	Ribonucleic acid
RRDR	Rifampicin Resistance-Determining Region
SAM	Sequence Alignment/Map format
SMRT	Single Molecule Real Time
SNP	Single Nucleotide Polymorphism
SRA	Short Read Archive
sSNP	Synonymous Single Nucleotide Polymorphism
SSTs	Second generation Sequencing Technologies
STB	Smooth Tubercle Bacillus
STR	Streptomycin
SV	Structural Variation
Syn	Synonymous
TB	Tuberculosis
TBDB	Tuberculosis Database
TDR	Total Drug Resistance
TDR-TB	Total Drug Resistance Tuberculosis
TIM	Targets of Independent Mutation

TSSs	Transcriptional Start Sites
VCF	Variant Call Format
WGS	Whole Genome Sequencing
WHO	World Health Organisation
XDR	Extensively Drug Resistance
XDR-TB	Extensively Drug Resistance Tuberculosis

Chapter 1

Introduction

1 INTRODUCTION

1.1 Tuberculosis disease and etiological agent

Tuberculosis (TB) is the second most common cause of death from an infectious disease worldwide, only behind the Human Immunodeficiency Virus (HIV) pandemic. According to the latest estimates from the World Health Organisation (WHO), there were 8.6 million new TB cases in 2012 and 1.3 million deaths, of which 0.3 million were HIV-associated (World Health Organization 2013). The majority of cases in 2012 occurred in Asia (58%) and Africa (27%), while smaller proportions occurred in the Eastern Mediterranean region (8%), Europe (4%) and the American continent (3%) (Figure 1.1). South-East Asia and Africa accounted for 75% of the total TB deaths. India and South Africa accounted for one third of TB deaths worldwide (World Health Organization 2013).

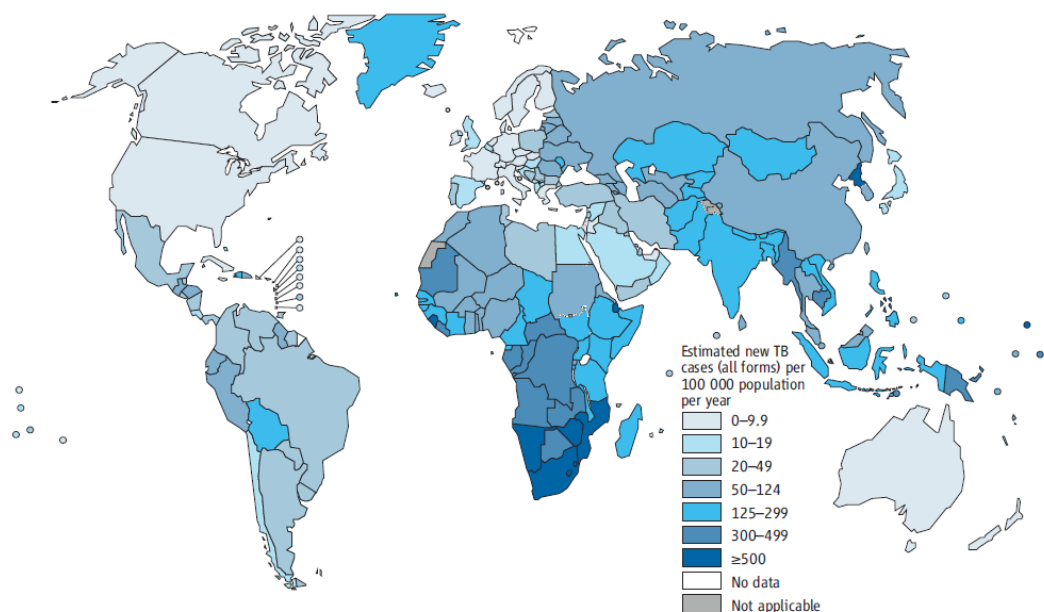
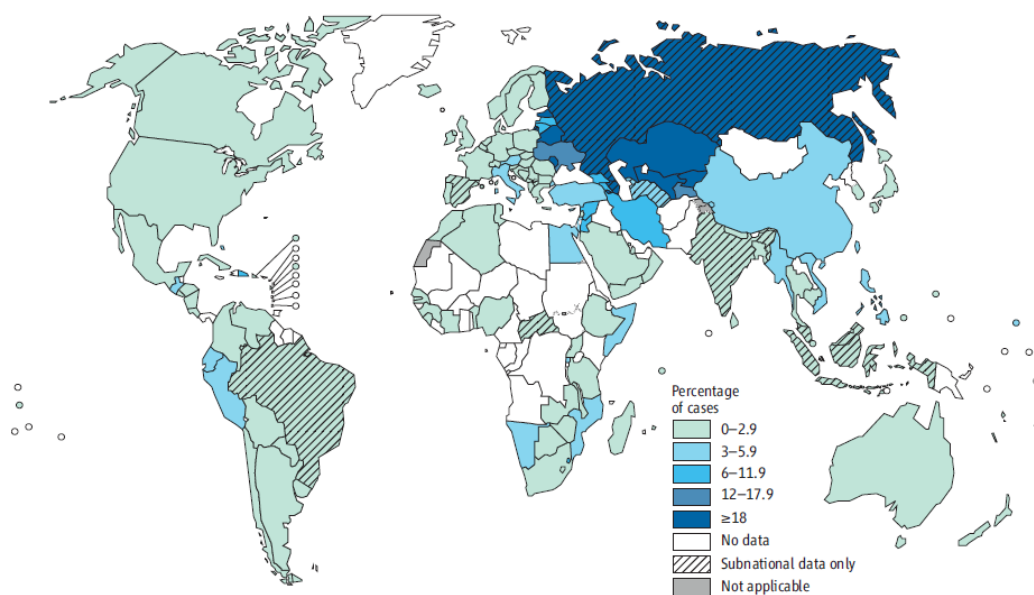


Figure 1.1 Estimated TB incidence rates in 2012

Figure reproduced from the Global Tuberculosis Report 2013(World Health Organization 2013)

Since 1990, estimated TB mortality rates have fallen by 45%. The WHO forecasts that a 50% decline in TB mortality will be achieved by 2015, compared to the baseline levels in 1990. Although incidence and mortality rates have been falling for the last decade due to the availability of efficacious treatments, TB remains a major global health threat. The emergence and spread of DR forms of TB further hinders TB control. The WHO classifies TB resistant to isoniazid (INH) and rifampicin (RMP), the two most effective first-line anti-TB drugs, as multi drug-resistant TB (MDR-TB). Globally, it is estimated that 3.6% of new TB cases and 20.2% of previously treated cases are MDR-TB (World Health Organization 2013). The highest levels of MDR-TB are reported in Eastern Europe and Central Asia (Figure 1.2). Globally, only a small proportion of TB cases are tested for MDR (5% of newly diagnosed TB cases and 9% of previously treated ones in 2012), of which not all cases are properly treated (World Health Organization 2013).



* Figures are based on the most recent year for which data have been reported, which varies among countries.

Figure 1.2 Percentage of new TB cases with MDR-TB in 2012
Figure reproduced from the Global Tuberculosis Report 2013(World Health Organization 2013)

Diagnosis and treatment of MDR-TB cases remain major challenges and are far from being fully achieved (Dheda *et al.* 2014). Extensively drug-resistance (XDR) strains, presumably emerged from MDR-TB strains and resistant to the second-line drugs fluoroquinolones (FLQ) and aminoglycosides (AMI), have been reported in 92 countries and, on average, 9.6% of MDR-TB cases are estimated to be XDR-TB (World Health Organization 2013).

Human TB is caused by bacteria belonging to the *M. tuberculosis* complex (MTBC), which is transmitted between people by inhalation of aerosol droplets that contain bacteria. TB cases are predominantly caused by *M. tuberculosis* (*Mtb*) followed by *M. bovis* and *M. africanum*, with occasional cases of infection with *M. caprae*, *M. microti*, *M. pinnipedii*, *M. orygis* and *M. canettii* reported. They are slow growing, lipid rich gram-positive actinomycetes with characteristic cell walls conferring natural resistance to many antibiotics (Brennan 2003). Members of the MTBC are indistinguishable in their 16SrRNA and *rpoB* genes, inter-strain recombination has not been reported and they have approximately the same genome length (Garcia-Betancur *et al.* 2011).

Once in the lungs, *Mtb* is phagocytised by macrophages, which are thought to be its preferential host cell during most of its life cycle, although this intracellular bacterium can infect different cell types in the host (Wolf *et al.* 2007; Randall *et al.* 2014). Uptake by macrophages triggers an initial innate immune response leading to the recruitment of inflammatory cells in the lungs (Cooper *et al.* 2011). *Mtb* is disseminated to the lymph nodes, where dendritic cells present bacterial antigens to naïve T-cells which then differentiate into antigen-specific effector T cells (Chackerian *et al.* 2002; Wolf *et*

al. 2008). Migration of these T cells to the infected lung stimulates the formation of granulomas which are composed of other cell types such as macrophages, lymphocytes and fibroblasts (Flynn *et al.* 2011) (Figure 1.3).

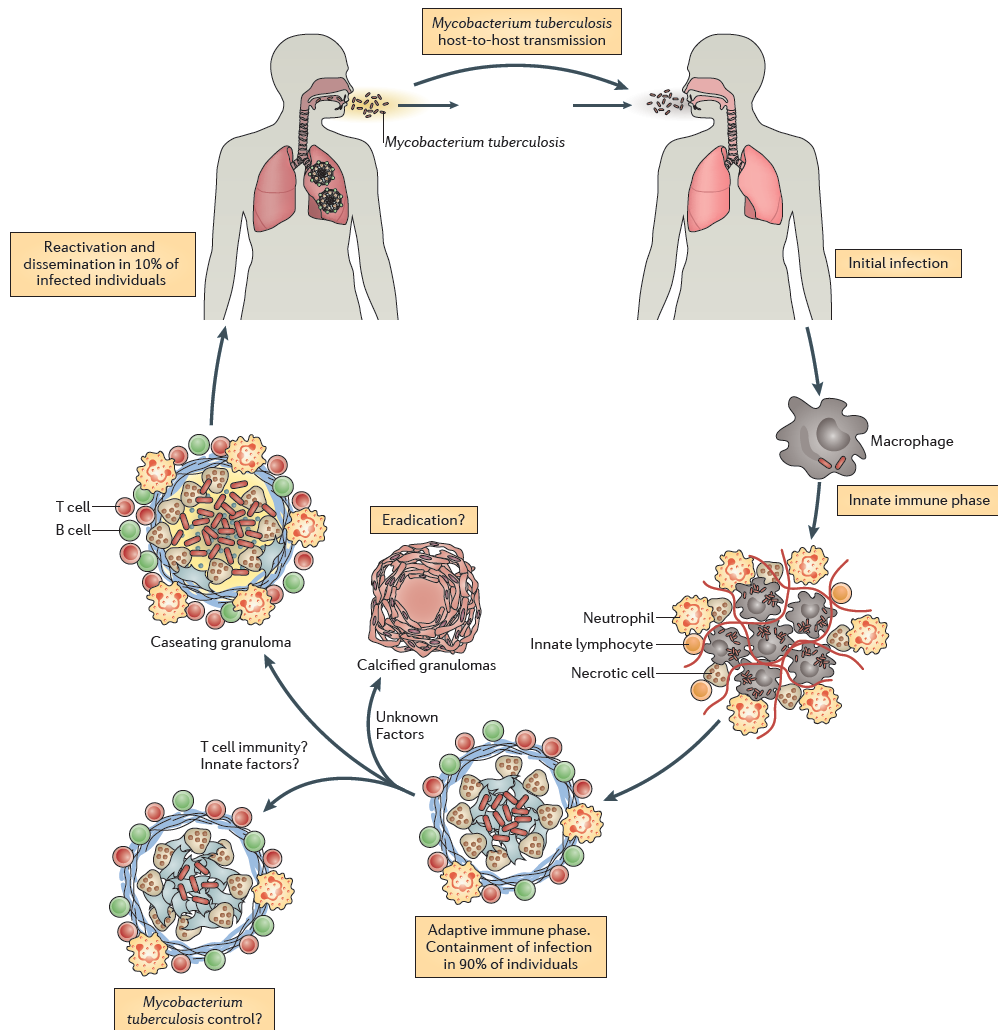


Figure 1.3 TB pathogenesis

Figure reproduced from (Nunes-Alves *et al.* 2014). Infection starts after inhalation of droplets containing *Mtb* bacteria. Phagocytosis of *Mtb* by macrophages triggers an innate immune response and recruitment of inflammatory cells. Recruitment of antigen-specific T cells and other immune cells leads to the formation of granulomas, where *Mtb* can remain latent.

Although this immune response controls the infection, *Mtb* can survive within the macrophages for months or even years in a clinically asymptomatic state, referred to as latent TB (Nunes-Alves *et al.* 2014). It has been estimated that one-third of the global human population may be latently infected by TB of which 5-10% may develop active disease sometime during their lives (Barry *et al.* 2009). Although most infection episodes will result in latent TB, the granuloma will break down in a small proportion of infected individuals, leading to active disease in which individuals become infectious and can transmit the infection (Barry *et al.* 2009).

Effective TB control relies on early diagnosis of the disease, access to treatment with anti-TB drugs and advances on vaccine development. Sputum smear microscopy has been the principal diagnostic test for TB during decades, and is still widely employed in low and middle income countries (Davis *et al.* 2013). Bacterial culture is the reference standard in TB diagnosis but results availability can take weeks due to the slow growth rate of MTBC bacteria. The rapid molecular test Xpert® MTB/RIF is considered an important breakthrough in TB diagnosis. This automated PCR assay can detect both MTBC bacteria and RMP resistance, normally within two hours and directly from sputum (Steingart *et al.* 2014). Xpert® MTB/RIF was endorsed by the WHO in 2010 and has been rapidly adopted by countries since then (World Health Organization 2013). Detection of resistance to anti-TB drugs relies on bacterial culture followed by drug susceptibility testing (DST). Solid culture generally takes between four to eight weeks and liquid culture, though more rapid than solid culture, still takes days and is more prone to contamination (Dheda *et al.* 2014).

Laboratory confirmation of TB and DR is crucial to ensure that infected people are given appropriate treatment. Standard regimens of first-line drugs applied for several months can usually eliminate TB. First-line therapies consist of combinations of drugs that were discovered more than 50 years ago and include INH (discovered in 1952), RMP (1966), ethambutol (EMB) (1961), pyrazinamide (PZA) (1952) and streptomycin (STR) (1943). MDR-TB strains are now widespread throughout the world, with about half a million cases reported in 2012 (World Health Organization 2013). To cure this type of strains, a switch to second line treatments is advised. Resistance to additional drugs such as EMB or STR further compromises treatment (Tahaoglu *et al.* 2001; Migliori *et al.* 2009). Second-line regimens use drugs such as FLQ and AMI, which are associated with multiple toxic effects and lower cure rates. These treatments have longer duration (for example, current regimens recommended by WHO entail at least 20 months) and may cost up to 100 times more than first-line treatments (Dheda *et al.* 2014). Increased resistance is associated with decreased patient survival and the emergence of resistance to first and second line drugs is a substantial threat to disease control. To date, resistance has been reported to all drugs used to treat TB (Dheda *et al.* 2014). New drugs are urgently needed to tackle the increasing problem of MDR and XDR-TB. Several new anti-TB drugs and regimens are currently under development (Zumla *et al.* 2014). Bedaquiline became the first new TB drug to be approved for use in 40 years (Andries *et al.* 2005).

An effective TB vaccine would be a powerful tool to eradicate TB. The use of Bacille Calmette-Guérin (BCG) (an attenuated form of *M. bovis*) has been widely implemented against human TB. However, its efficacy is variable between human populations and

confers low protection in developing countries (Fine 1995). More than 12 candidate vaccines are currently being tested in clinical trials (Ottenhoff & Kaufmann 2012). These vaccines aim to induce T cell-mediated immunity required to prevent progression into active pulmonary disease (Nunes-Alves *et al.* 2014) and ultimately stop transmission, either alone or following BCG vaccination. The current understanding of protective immunity against *Mtb* after infection is incomplete, which hinders vaccine development (Kaufmann *et al.* 2014).

1.2 Impact of WGS on TB research and clinical applications

There is an urgent need for better treatments and vaccines, which in turn require a deeper understanding of the biology of *Mtb*. Knowledge of the genomic variability among *Mtb* isolates could result in such biological insights (Comas & Gagneux 2009), given the increasing evidence that strain genetics may play a role in disease outcome, transmission, variation in vaccine efficacy (López *et al.* 2003) or emergence of DR (Ford *et al.* 2013).

The application of genomics to the study of TB has greatly improved our understanding of this disease. Thanks to the still growing use of genome sequencing and comparative genomics, the TB research community has gained new insights into the origins (Galagan 2014; Comas *et al.* 2013), within-host microevolution (Pérez-Lago *et al.* 2014; Casali & Nikolayevskyy 2012), epidemiology (Pérez-Lago *et al.* 2014; Bryant, Harris, *et al.* 2013; Walker *et al.* 2013; Gardy *et al.* 2011) and DR genetic determinants of *Mtb* (Casali *et al.* 2014).

The first *Mtb* genome to be fully sequenced was that of the H37Rv laboratory strain in 1998 (Cole *et al.* 1998). The complete genome sequence was determined in order to improve the understanding of *Mtb* biology and aid in the development of new therapies and vaccines. The circular genome comprising 4.4 million base pairs (Mb) was found to contain around 4,000 genes and to have a relatively high GC content (65%). Unlike other bacterial genomes, *Mtb* genome was found to encode a large number of enzymes involved in lipid metabolism. Some of these produce multiple and diverse lipophilic molecules, ranging from simple fatty acids to very-long-chain and highly complex mycolic acids, the predominant lipid component of the mycobacterial cell wall (Brennan 2003). Other enzymes are used to degrade host-cell lipids, particularly fatty acids and cholesterol, and used as energy sources during intracellular growth and persistence (Ouellet *et al.* 2011). Two protein families (PE and PPE genes) were found to comprise about 10% of the coding potential of the genome. These proteins have a repetitive structure, are highly polymorphic and their function remains largely unknown. Although initially suggested to be involved in antigenic variation, their possible function as variable surface antigens and their role in immune evasion is still an area of active research (Copin *et al.* 2014; Comas *et al.* 2010). Since its publication in 1998, the H37Rv reference genome has been functionally annotated with information from the scientific literature (Lew *et al.* 2011) and nowadays contains a total of 4,018 protein genes, 13 pseudogenes and 80 RNA loci.

Whole-genome sequencing (WGS) of multiple strains has also enabled more reliable reconstructions of the phylogenetic history of MTBC (Comas *et al.* 2013). Both archaeological findings and comparative genome analyses conflict with a zoonotic

origin of human-adapted TB. According to the zoonotic hypothesis, an ancient *M. bovis* strain could have been transferred from cows to humans during animal domestications.

A new scenario of the evolutionary history of MTBC has emerged. There is mounting evidence that MTBC originated from a common ancestor in the Horn of Africa, most likely from a smooth tubercle bacillus (STB) like *M. canettii* (Figure 1.4).

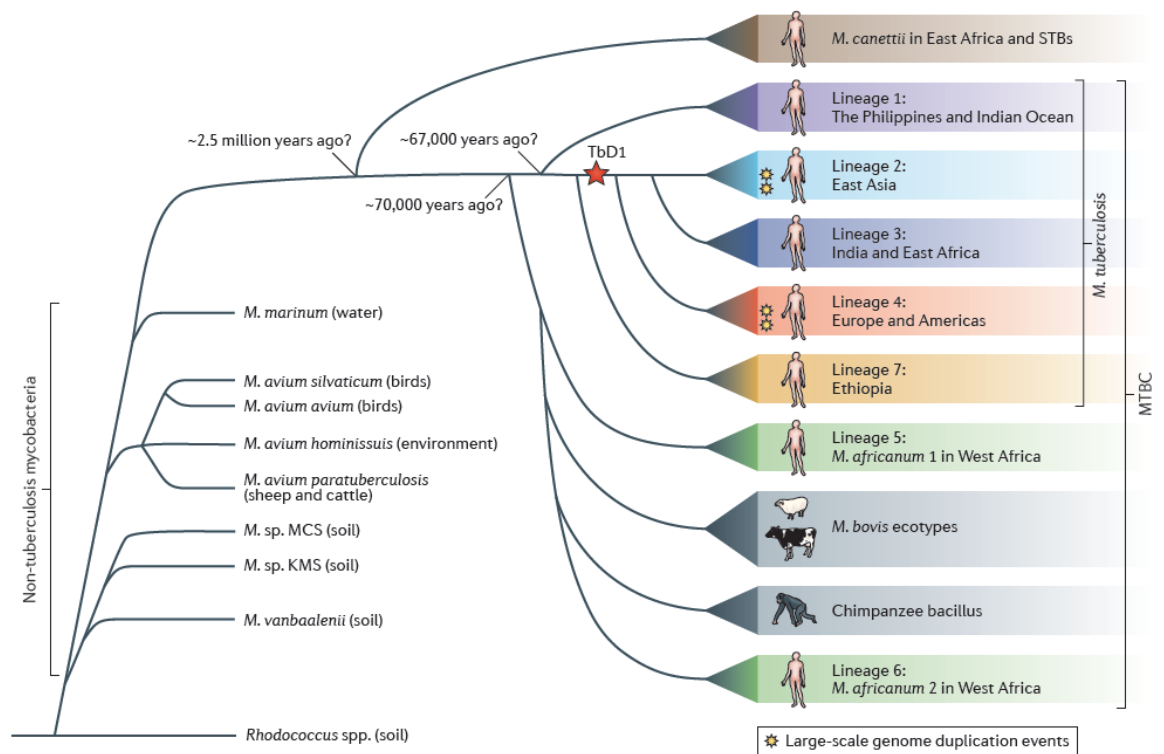


Figure 1.4 MTBC evolutionary history

Figure reproduced from (Galagan 2014).

While STB strains are genetically diverse and show evidence of extensive recombination and horizontal gene transfer (HGT), MTBC seems to have arisen as a clonal expansion from a single STB progenitor (Supply *et al.* 2013), and displays low

genetic diversity as a consequence of this clonal population structure. After MTBC emerged in Africa, it spread with the first human migrations (Blouin *et al.* 2012; Comas *et al.* 2013; Supply *et al.* 2013) and differentiated into six major global lineages associated with a restricted geographical area (Gagneux *et al.* 2006).

In addition to the study of macroevolutionary episodes of TB, WGS has also been applied to investigate the microevolution events occurring within individual patients and between different patients along transmission chains. WGS can differentiate between relapse and re-infection in cases of recurrent TB with greater resolution and accuracy than standard fingerprinting techniques (RFLP and MIRU-VNTR) (Bryant, Harris, *et al.* 2013). WGS has proved to be successful in delineating community outbreaks of TB, inferring chains of transmission between cases and identifying super-spreaders (Gardy *et al.* 2011; Walker *et al.* 2014; Walker *et al.* 2013; Roetzer *et al.* 2013), i.e. particularly infectious individuals leading to many secondary cases. The genetic determinants of recurrence and transmissibility can additionally be examined using WGS and gain insight into their biological basis. The mutation rate (also referred to as substitution or evolutionary rate) can be determined using longitudinal samples from the same patient (Walker *et al.* 2013; Bryant, Harris, *et al.* 2013) or dated samples from different individuals of the same outbreak (Roetzer *et al.* 2013). All these studies reported similar estimates of the substitution rate, 0.3-0.5 mutations per genome per year. This mutation rate is 10-fold lower than that of methicillin-resistant *Staphylococcus aureus* (Nübel *et al.* 2010) and explains, in part, the limited sequence diversity of MTBC (Schürch *et al.* 2010). The substitution rate is a valuable measure as it can be used to estimate the common ancestor originating date of outbreak-

circulating strains (Roetzer *et al.* 2013). The choice of a minimum number of SNPs to establish epidemiological links between TB infected patients has been more controversial. The recent finding that the genetic diversity accumulated within a patient can be as high as that observed between patients (Pérez-Lago *et al.* 2014) poses challenges for establishing thresholds and inferring transmission events.

The application of WGS to the study and diagnosis of DR TB holds great promise (Rodwell *et al.* 2013; Sharon J. Peacock 2013; García-Sierra *et al.* 2011; Lin *et al.* 2013). Efforts to reduce the prevalence of DR TB are focused on rapid detection of DR cases, effective treatment of those and prevention of ongoing transmission. These activities rely on antimicrobial susceptibility testing and bacterial genotyping, which may take several months to be accomplished because of the slow growth rate of MTBC. In this context, WGS has the potential of being applied as a tool for high-discriminatory genotyping and DR diagnosis. The DR mutations in patient-isolated *Mtb* strains can be used to predict which drugs might be more clinically effective for a particular patient. Cases of developed resistance, i.e. arisen spontaneously during drug treatment, could also be distinguished from cases of transmitted resistance, i.e. due to re-infection with a resistant strain (Clark *et al.* 2013), based on the presence of phylogenetically informative mutations. However, WGS cannot replace phenotypic susceptibility testing for all antibiotics, given the incomplete understanding of the genetic causes of resistance (Sharon J. Peacock 2013) of some of them (Bhujji *et al.* 2013; Brossier *et al.* 2011). In this regard, WGS can be a powerful research tool to dissect the genetic determinants of antibiotic resistance (H. Zhang *et al.* 2013; Farhat *et al.* 2013).

1.3 Next-generation sequencing technologies

Second generation sequencing technologies (SSTs), also refer as to next-generation sequencing (NGS), were designed to parallelise the sequencing process to deliver high-throughput sequences at lower cost than standard capillary sequencing (Sboner *et al.* 2011). Table 1.1 summarises the main features of the most established SSTs platforms compared to traditional Sanger capillary sequencing (Liu *et al.* 2012; Glenn 2011).

Table 1.1 Commercially available second generation sequencing platforms

	Illumina		SOLID	454		Sanger
	Illumina GAIIx	Illumina HiSeq2000	SOLiDv4	GS Junior System	GS FLX+ System	3730xl
Sequencing method	Sequencing by synthesis		Ligation and 2-base coding	Pyrosequencing		Dideoxy chain termination
Maximum Output	95 Gb	600 Gb	120 Gb	~35 Mb	~700 Mb	1.9~84Kb
Maximum read length	2 x 100	2 x 150	50 + 35	400 bp	700 bp	900 bp
Millions of reads per run	320	3000	840	0.1	1	0.000096
Run time	14 days	8 days	12 days	10 hours	23 hours	2h
Cost per Mb	\$0.12	\$0.10	\$0.11	\$22	\$10	\$1500
Instrument cost	\$250,000	\$690,000	\$475,000	\$108,000	\$500,000	\$376,000
Main advantage	High throughput		Low error rates	Read length, fast		High quality, long read length
Main disadvantage	Short read assembly		Short read assembly	High cost, low throughput		High cost low throughput
Primary applications	Transcriptome characterization, de novo large genomes, re-sequencing, transcript counting, mutation detection, and metagenomics		Re-sequencing, transcript counting and mutation detection	De novo microbial genomes, transcriptome characterization and metagenomics		De novo microbial and large genomes, mutation detection

Current Illumina sequencers, e.g. Illumina Genome Analyser II and HiSeq2000, can sequence hundreds of bacterial genomes of a few mega bases (Mb) to at least 50-fold coverage in a single run (www.illumina.com). Because of the short length of sequencing reads (50-250 bp), these platforms have been extensively employed in re-sequencing projects, i.e. when a complete genome from a close strain or species is already available. Despite differences in read lengths, depth of coverage (DOC) and other features there is an increasing overlap for the same applications among different platforms (Table 1.1). A broad spectrum of bioinformatic algorithms has flourished to meet the needs of sequence data analysis, management and interpretation.

1.4 Whole genome sequence analyses

Current high throughput sequencing machines produce tens of millions of sequences in a single run. Raw sequence data is generally stored in files of *FASTQ* format (Cock *et al.* 2010), a text-based format for storing both the nucleotide sequence and its corresponding quality scores. Quality control and filtering are firstly applied to raw reads in order to minimise the artefacts arising during the sequencing reactions, including base calling errors, poor quality reads or primer contamination. After removing poor quality reads and samples, sequenced reads are typically mapped to a reference genome.

The first step in most sequence analysis pipelines involves the mapping or alignment of reads against a reference genome, a requisite stage for downstream analyses. With the introduction of NGS platforms, traditional alignment software became obsolete

In addition to the mapping approach, *de novo* assembly can be used to reconstruct the target sequence not assisted by the comparison to previously resolved reference sequences. It follows a bottom-up strategy by which reads are grouped into contigs and those into scaffolds covering, ideally, the whole chromosome length. Despite the limitations initially imposed by NGS data, high coverage currently achieved (e.g. 100x by Illumina HiSeq2000), growing read lengths (150 bp by Illumina HiSeq2000) and paired-end information makes it feasible to obtain relatively low fragmented assemblies from bacterial genomes (Magoc *et al.* 2013). *De Bruijn* graph assemblers, such as *Velvet* (Zerbino & Birney 2008) or *SOAPdenovo*, are among the most commonly used and have become the programs of choice when processing short reads produced by Illumina and SOLiD platforms (25-150 bp range) (W. Zhang *et al.* 2011).

Most of software tools for variant detection and calling require as input alignment files obtained by mapping software, commonly in *SAM/BAM* format. Single nucleotide polymorphisms (SNPs) can be distinguished from sequencing errors thanks to the high DOC achieved by SSTs. True SNPs are expected to occur as mismatches across multiple reads at the same reference position whereas mismatches found at spurious locations are likely to be sequencing errors (Figure 1.6). SNP calling tools (Nielsen *et al.* 2011) make use of this information to calculate statistical significance and filter out false positive SNPs. Small indels, namely those shorter than the read length, can also be called since mapping algorithms allow for gapped alignments.

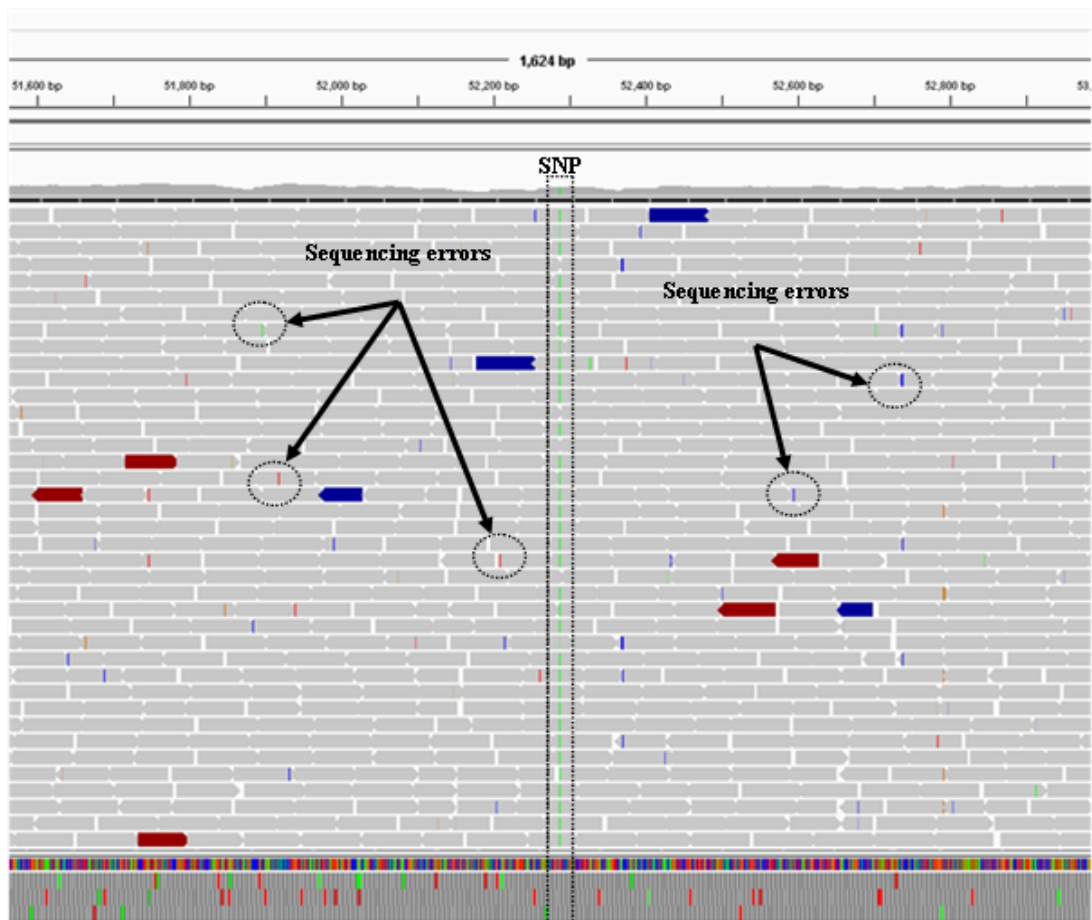


Figure 1.6 Alignment visualisation with a SNP and sequencing errors

Alignment of sequencing reads to a region of the *Mtb* reference genome. Reads are displayed as rectangles in gray if matching the reference sequence. Mismatches are colour-coded by nucleotide: A green, C blue, G yellow and T red. A mismatch is found across most of the reads centred in the figure (probable SNP). Spurious mismatches are likely to be sequencing errors.

Structural variation (SV) refers to relatively large polymorphisms that alter the chromosome structure (e.g. indels, inversions and copy number variants (CNV)) (Alkan *et al.* 2011). New tools have been developed to detect unusual patterns of reads, or pairs of reads, left by structural variants. Such signatures can be broadly grouped into three categories: signatures based on discordant mapping of read pairs, signatures based on read splitting and signatures based on DOC (Alkan *et al.* 2011). SV programs

implement algorithms aiming to identify such signatures, or combinations of them, from sequence alignment files (Layer *et al.* 2014; Rausch *et al.* 2012).

The most powerful variation calling approach would be to perform *de novo* assembly of each genome and identify polymorphisms by alignment of genome sequences. This would enable unbiased detection of all types and lengths of SVs. It is expected that *de novo* assembly of genomes followed by subsequent pair-wise comparison to the reference genome will become the standard method of SV detection. This approach has not been reliable to resolve all genomic regions, especially regions with repeats and duplications, due to limited read lengths and differences in coverage between regions. Nevertheless, with increasing read lengths and DOC obtained by the latest generation of sequencers it will be possible to produce relatively low fragmented assemblies from bacterial genomes (Utturkar *et al.* 2014).

1.5 Research aims and objectives

The overall aim of this work is to address the bioinformatic challenges that are associated with the analysis and interpretation of WGS data derived from *Mtb* clinical isolates and advance towards a more complete understanding of the genomic diversity of *Mtb*. Each of the specific objectives are summarised in the following points:

1. Design and implementation of bioinformatic pipelines to derive genomic variants from MTBC raw sequence data, making use of the state-of-the-art mapping and *de novo* assembly bioinformatic tools.

2. Development of an open-access web-based resource of MTBC genetic polymorphisms derived from publicly-available WGS projects.
3. Development of *in silico* genotyping approaches to bridge the gap between classical genotyping and high throughput sequencing.
4. Define a set of lineage and sub-lineage specific markers that can be used to discriminate known circulating strains, both accurately and robustly.
5. Study the diagnostic performance of known DR mutations as markers for predicting phenotypic resistance from WGS data.
6. Discovery of new genes involved in DR.

In summary, the objectives of this work involve developing bioinformatic tools for processing and making MTBC genomic data accessible, as well as identifying informative genetic markers, both strain-specific and DR-associated, to barcode MTBC strains in the context of epidemiological, diagnostic and clinical studies.

1.6 Description of the thesis and contributions

The content of this thesis corresponds to that of five research papers (Table 1.2), in addition to the Introduction (Chapter 1) and Discussion and Further Work (Chapter 6), which are produced specifically for this thesis. Chapters 2, 4 and 5 are composed of one research paper each, while Chapter 3 comprises the content of two research papers (Table 1.2).

Table 1.2 Research papers included in this thesis in chronological order

Research Paper Number (Chapter)	Authors	Title	Status, journal and year
1 (Chapter 3)	Francesc Coll, Kim Mallard, Mark D. Preston, Stephen Bentley, Julian Parkhill, Ruth McNerney, Nigel Martin and Taane G. Clark	SpolPred: rapid and accurate prediction of <i>Mycobacterium tuberculosis</i> spoligotypes from short genomic sequences.	Published. Bioinformatics (2012)
2 (Chapter 2)	Francesc Coll, Mark D. Preston, José Afonso Guerra-Assunção, Grant Hill-Cawthorn, David Harris, João Perdigão, Miguel Viveiros, Isabel Portugal, Francis Drobniowski, Sebastien Gagneux, Judith R. Glynn, Arnab Pain, Julian Parkhill, Ruth McNerney, Nigel Martin and Taane G. Clark	PolyTB: A genomic variation map for <i>Mycobacterium tuberculosis</i> .	Published. Tuberculosis (2014)
3 (Chapter 3)	Francesc Coll, Ruth McNerney, José Afonso Guerra-Assunção, Judith R Glynn, João Perdigão, Miguel Viveiros, Isabel Portugal, Arnab Pain, Nigel Martin and Taane G Clark.	A robust SNP barcode for typing <i>Mycobacterium tuberculosis</i> complex strains.	Published. Nature communications (2014)
4 (Chapter 4)	Francesc Coll, Ruth McNerney, Mark D Preston, José Afonso Guerra-Assunção, Andrew Warry, Grant Hill-Cawthorne, Kim Mallard, Mridul Nair, Anabela Miranda, João Perdigão, Miguel Viveiros, Isabel Portugal, Zahra Hasan, Rumina Hasan, Judith R Glynn, Nigel Martin, Arnab Pain and Taane G Clark.	Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences	Submitted
5 (Chapter 5)	Francesc Coll, Grant A. Hill-Cawthorne, Kim Mallard, Rumina Hasan, Zahra Hasan, Nerges Mistry, Rob Warren, Keertan Dheda, Patricia Sheen, David Moore, Jaime Robledo, Maxine Caws, Stefan Pantaiotov, Richard Anthony, Saad Alghamdi, Joao Perdigao, Miguel Viveiros, Isabel Portugal, Andy Ramsey, Bouke de Jong, Leen Rigouts, Theolis Bessa, Tomoshige Matsumoto, Anabela Miranda, Noram Mocillo, Christophe Sola, Ruth McNerney, Arnab Pain and Taane G Clark.	A whole genome association approach reveals insights into global <i>Mycobacterium tuberculosis</i> drug resistance	In preparation

Chapter 2 corresponds to Research Paper 2, titled ‘PolyTB: A genomic variation map for *Mycobacterium tuberculosis*’. Chapter 2 describes the implementation of the

genome variation discovery pipeline (Research Paper 2). It is applied across independent WGS data sets - sourced from epidemiological, DR and evolutionary studies - as a first step for phylogenetic, population genetic and other downstream analyses. *PolyTB* is also presented, a repository hosting the discovered genomic variants, annotated and integrated with strain type and geographical metadata. TGC and I conceived and designed this study, which was jointly supervised by TGC and NM. I developed and tested the genomic discovery bioinformatic pipelines with input from JAG, MDP and TGC. I developed *PolyTB* using as a starting point *PlasmoView* - a project coded by MDP to display malaria genomic data (Preston *et al.* 2014) - with input from all other authors and technical advice from MDP, TGC and NM. I drafted, wrote and finalised the manuscript with contributions from all other authors, and produced all tables and figures in the manuscript, including all summary statistics and phylogenetic analyses. DH, JP, MV, IP, FD, SG, JRG, AP, JP and RM contributed to the sequencing of samples and metadata. Research Paper 2 was published in *Tuberculosis* on the 8th of February 2014.

Chapter 3 is composed of two research papers, research papers 1 and 3 (Table 1.2), both related to *Mtb* typing from whole genome sequences. Approaches that can predict traditional genotypes from WGS data are investigated, leading to the development of *SpolPred*, a software tool for accurate determination of *Mtb* spoligotype patterns from WGS data (Research Paper 1). TGC and I conceived and designed this study. I developed and tested *SpolPred* software with the contribution of MDP, who helped me optimise the performance and speed of the tool. KM carried out the experimental spoligotyping, whose results were compared to *SpolPred* predicted

ones. TGC, NM and RM jointly supervised the project. RM provided the samples and SB and JP contributed to the sequencing of these. RM, TGC and I drafted and finalised the manuscript with contributions from all other authors. Research Paper 1 was published in *Bioinformatics* on the 29th of August 2012.

Given the limitations of traditional genotyping techniques and the growing consensus on the use of SNPs as robust and highly discriminatory phylogenetic markers (Comas *et al.* 2009), a new SNP-based classification system for MTBC strains is proposed in Chapter 3 (Research Paper 3). TGC and I designed this study. JAG, JRG, JP, MV, IP, and AP contributed to the construction of the data set. RN, NM and TGC jointly supervised the research. I conducted all analyses, including the bioinformatic pipelines to derive high quality genomic variants and phylogenetic and population genetic analyses, and produced all tables and figures. RM, TGC and I wrote the paper with contributions from all other authors. Research Paper 3 was published in *Nature Communications* on the 1st of September 2014.

Chapter 4 corresponds to Research Paper 4 titled 'Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences'. To assess the potential benefits of a whole genome approach to detect DR TB, an updated library of mutations predictive of DR has been curated from the literature. *TB profiler*, an online tool for analyzing raw sequence data and predicting resistance, has been implemented. The prediction of strain type, based on strain-specific mutations, has been added to enhance the usefulness of this tool. RM, AP, TGC and I conceived and designed the study. MDP and I developed and tested the online tool *TB profiler*; JAG and AW

developed additional software which were not core to the final work. GH-C, MN and KM performed laboratory experiments and curation of meta data for sequencing. AM, JP, MV, IP, ZH, RH, JRG contributed biological samples, sequencing or phenotypic data. I searched the literature and curated the database of DR-associated mutations. I conducted the bioinformatic analyses to derive high quality genomic variants, performed comparisons of experimentally determined phenotypes with predicted ones, and performed the statistical analysis under the guidance of NM and TGC. I produced all tables and figures in the manuscript. AP led the sequencing efforts. RM, TGC and I drafted, wrote and finalised the manuscript with contributions from all other authors. Research Paper 4 was submitted for publication to *Genome Biology* on December 2014.

Chapter 5 corresponds to Research Paper 5, in which phenotype-genotype association analyses are performed and novel loci associated with resistance identified. I conducted the bioinformatic analyses required to obtain a high quality data set of genomic variants, and performed the phylogenetic and genome-wide association analyses under the guidance of TGC. GA-CH, KM, and AP coordinated the sequencing. RH, ZH, NM, RW, KD, PS, DM, JR, MC, SP, RA, SA, JP, MV, IP, AR, BdJ, LR, TB, TM, AM, NM and CS contributed DNA samples and meta data, including strain-typing and drug susceptibility testing data. RM, AP, and TGC are joint PIs on the project.

The final Chapter 'Discussion and Further Work' does not correspond to a research paper, but summarises the main findings, places the research into a wider TB context, discusses the limitations of the thesis and outlines opportunities for future research.

1.7 Description of the data sets

Multiple MTBC populations from independent studies have been used throughout this work. Table 1.3 provides a brief description of each of these studies. The population name is given based on the geographical source of the samples. Only two studies consisted of samples from multiple sources: the Global key strains (Comas *et al.* 2013) and WHO-TDR studies (Vincent *et al.* 2012). For studies with publicly-available WGS data, the European Nucleotide Archive (ENA) accession number is provided. The sequencing centre, sequencing technology, sample size, read length and median DOC of each study are also provided. The availability of phenotypic DR testing is additionally indicated. Table 1.4 describes the composition of each data set used in the chapters, namely which studies they are composed of.

Table 1.3 Summary of all WGS TB studies

Population (reference)	ENA accession number	Sample Size	Read length	DOC	DR available
Samara, Russia (Casali & Nikolayevskyy 2012)	ERP000192 ^c	329	49 ^b	61	Yes (42/329)
Midlands, UK (Walker <i>et al.</i> 2013)	ERP000276 ^c	390	75 ^b	112	No
Kampala, Uganda (Clark <i>et al.</i> 2013)	ERP000520 ^c	51	75 ^a	257	Yes
Global key strains (Comas <i>et al.</i> 2013)	ERP001731 ^c	171	75/100 ^b	97	No
Bilthoven, Netherlands (Bryant, Schürch, <i>et al.</i> 2013)	ERP000111 ^c	213	75/100 ^b	39	No
Vancouver, Canada (Gardy <i>et al.</i> 2011)	SRP002589 ^d	36	50 ^b	37.5	Yes
Lisbon, Portugal (Perdigão <i>et al.</i> 2013)	ERP002611 ^e	84	100 ^a	104	Yes
Karonga, Malawi (A) (Guerra-Assunção <i>et al.</i> 2014)*	ERP000436 ^c	353	75 ^a	183	Yes
Karonga, Malawi (B) (Guerra-Assunção <i>et al.</i> 2014)	ERP000436 ^c	1662	75 ^a	102	Yes
China (H. Zhang <i>et al.</i> 2013)	SRA065095 ^f	161	75 ^a	113	Yes
Djibouti (Blouin <i>et al.</i> 2012)	ERP001885 ^g	7	75 ^a	75	No
Ethiopia (Firdessa <i>et al.</i> 2013)	ERP001567 ^h	4	75 ^a	95	No

Porto, Portugal	- ^e	128	100 ^b	157	Yes
Karachi, Pakistan	- ^e	42	100 ^b	448	Yes
Brazil	- ^e	108	100 ^b	110	Yes
Bulgaria	- ^e	17	100 ^b	155	No
Colombia	- ^e	15	100 ^b	111	Yes
India	- ^e	17	100 ^b	59	No
Japan	- ^e	4	100 ^b	283	No
Netherlands	- ^e	14	100 ^b	389	Yes
Peru	- ^e	104	100 ^b	290	Yes
South Africa	- ^e	174	100 ^b	188	Yes
Vietnam	- ^e	50	100 ^b	192	Yes
WHO-TDR	- ^e	190	100 ^b	123	No

^aSequenced using Illumina HiSeq2000, ^bSequenced using Illumina Genome Analyzer II, ^cSequenced at the Wellcome Trust Sanger Institute, ^dSequenced at Simon Fraser University, ^eSequenced at the King Abdullah University of Science and Technology (KAUST), ^fSequenced at the Beijing Genomics Institute (BGI), ^gSequenced at the Institut de Génétique et Microbiologie, Université Paris Sud, ^hSequenced at the Center for Public Health Research, University of Valencia.*Karonga, Malawi (A) is a sub-set of Karonga, Malawi (B)

Table 1.4 Composition of the four WGS data sets

Data set name	Populations included	
WGS data set 1	Samara, Russia; Midlands, UK; Kampala, Uganda; Global key strains; Bilthoven, Netherlands; Vancouver, Canada; Lisbon, Portugal; Karonga, Malawi.	Chapter 2
WGS data set 2	WGS data set 1 + China, Djibouti and Ethiopia.	Chapter 3
WGS data set 3	China; Karachi, Pakistan; Karonga, Malawi; Lisbon, Portugal; Porto, Portugal; and Samara, Russia.	Chapter 4
WGS data set 4	Brazil; Bulgaria; China; Colombia; Vancouver, Canada; India; Japan; Karachi; Karonga, Malawi; Lisbon, Portugal; Netherlands; Peru; Porto; Samara, Russia; South Africa; Kampala, Uganda; Vietnam; WHO-TDR	Chapter 5



Registry

T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

COVER SHEET FOR EACH 'RESEARCH PAPER' INCLUDED IN A RESEARCH THESIS

Please be aware that one cover sheet must be completed for each 'Research Paper' included in a thesis.

1. For a 'research paper' already published

1.1. Where was the work published? Tuberculosis (Edinburgh, Scotland).....

1.2. When was the work published? 08/02/2014.....

1.2.1. If the work was published prior to registration for your research degree, give a brief rationale for its inclusion

.....
.....
.....

1.3. Was the work subject to academic peer review? Yes.....

1.4. Have you retained the copyright for the work? **Yes / No**

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 copyright holder (publisher or other author) to include work

2. For a 'research paper' prepared for publication but not yet published

2.1. Where is the work intended to be published?

2.2. Please list the paper's authors in the intended authorship order

.....

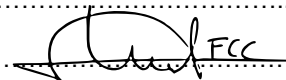
2.3. Stage of publication – Not yet submitted / Submitted / Undergoing revision from peer reviewers' comments / In press

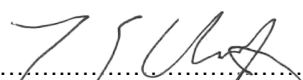
3. For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)

.....
See next page
.....

NAME IN FULL (Block Capitals) FRANCESC COLL I CEREZO.....

STUDENT ID NO: 323873.....

CANDIDATE'S SIGNATURE ..... **Date** 10/12/2014.....

SUPERVISOR/SENIOR AUTHOR'S SIGNATURE (3 above) .....

RESEARCH PAPER 2

PAPER DETAILS:

Francesc Coll, Mark D. Preston, José Afonso Guerra-Assunção, Grant Hill-Cawthorn, David Harris, João Perdigão, Miguel Viveiros, Isabel Portugal, Francis Drobniowski, Sebastien Gagneux, Judith R. Glynn, Arnab Pain, Julian Parkhill, Ruth McNerney, Nigel Martin, Taane G. Clark (2014). PolyTB: A genomic variation map for Mycobacterium tuberculosis. *Tuberculosis (Edinburgh, Scotland)*, 1–9. doi:10.1016/j.tube.2014.02.005

AUTHORS CONTRIBUTION:

FC and TGC conceived and designed the study, which was jointly supervised by TGC and NM. FC developed and tested the genomic discovery bioinformatic pipelines with input from JAG, MDP and TGC. FC developed PolyTB using as a starting point PlasmoView - a project coded by MDP to display malaria genomic data - with input from all other authors and technical advice from MDP and NM. FC drafted and finalised the manuscript with contributions from all other authors. FC produced all tables and figures in the manuscript, including all summary statistics, and performed phylogenetic analyses. DH, JP, MV, IP, FD, SG, JRG, AP, JP and RM contributed to the sequencing all samples and metadata of those. The final version of the manuscript was read and approved by all authors.

Chapter 2

Identifying and Visualising Genomic Variation

2 IDENTIFYING AND VISUALISING GENOMIC VARIATION

An increasing number of WGS samples have become publicly available during the past few years - sourced from epidemiological, DR and evolutionary studies in *Mtb* - and deposited as raw sequence files. The aim of the work presented in this chapter is to process this wealth of data, mine the genetic variation in it and present it to the TB community in an integrated and intuitive manner.

2.1 INTRODUCTION

SNPs, indels and other genetic polymorphisms derived from WGS provide enough discriminatory power to assess natural variation in populations. These include variants associated with the host-pathogen relationship, including virulence factors, drug susceptibility determinants and immune modulator factors with importance on the clinical manifestations (Ford *et al.* 2012). Due to the low mutation rate (Bryant, Schürch, *et al.* 2013) and limited genomic diversity of MTBC, the application of WGS in clinical settings is particularly effective for *Mtb* (Köser *et al.* 2012).

After the first *Mtb* genome was sequenced in 1998 (Cole *et al.* 1998), another 26 complete 'reference' genomes have been sequenced and made publicly available (NCBI 2014). Databases like the Mycobacterial Genome Divergence Database (MGDD) (Vishnoi *et al.* 2008), the Single Nucleotide Polymorphism Database (dbSNP) (Smigielski *et al.* 2000), and Tuberculosis Database (TBDB) (Reddy *et al.* 2009) curate genomic variants across some of these available complete genomes. The Pathosystems Resource Integration Center (PATRIC) houses genomes of different bacterial pathogens

including MTBC strains (Wattam *et al.* 2014), 449 to date. In terms of gene annotation, *Tuberculist* provides exhaustive and updated functional information such as operon annotation or protein information (Lew *et al.* 2011).

The TB community has a number of available web-based databases and tools to exploit the existing molecular epidemiological data (Shabbeer *et al.* 2012), SNP repositories (Stucki & Gagneux 2012) and manually-annotated genomes (Sandgren *et al.* 2009). Nevertheless, there is no tool harbouring genetic polymorphisms derived from WGS projects integrated with geographic distribution, strain type information and population structure visualisation. Despite the number of WGS data sets stored in public repositories like the ENA or SRA (Short Read Archive), users cannot browse, compare and contextualise the genomic variants resulting from these studies. Clinicians, epidemiologists and researchers working on TB would benefit from a resource allowing the investigation of genetic variation at genes of interest and geographic distribution of strains and clinically important genetic variants such as DR markers.

To fill this gap, *PolyTB* was developed, a web-based tool to display MTBC genetic polymorphisms derived from publicly available WGS datasets. A catalogue of SNPs, small indels and large deletions was compiled by employing the state-of-the-art variation discovery software (Alkan *et al.* 2011). Variants can be investigated through a genome browser reporting their chromosome coordinates, and a world map showing their global allele distribution. Additionally, the construction of phylogenetic trees based on SNPs provides an additional tool to investigate the population structure.

Strain genotype information is incorporated, allowing the visualisation of associations of strain types with particular polymorphisms and/or geographical locations as well as aiding correlation with public health epidemiological data. The integration of such data into tools like *PolyTB* is required to fully exploit genomic variation, and potentially boost TB control research through the discovery of new drug targets, vaccine antigens and diagnostics.

2.2 METHODS

2.2.1 Genomic variation discovery pipeline from Illumina paired-end sequence data

MTBC isolates from the WGS dataset 1 described in Section 1.7 were downloaded from the ENA (<http://www.ebi.ac.uk/ena/>). All isolates (n = 1,627) had been sequenced using Illumina paired-end technology (Illumina-GAll or HiSeq 2000). For each of the samples, *Trimmomatic* software version 0.27 (Lohse *et al.*, 2012) was used to clean the raw data, removing low quality reads and low-quality 3' ends of reads, and keeping only reads at least 36 base pairs long, with nucleotides above Q20. Filtered sequences were then aligned to the H37Rv reference genome (Genbank accession number: NC_000962.3) using the *BWA mem* algorithm (version 0.7.9a-r786) (Langmead *et al.* 2009). Default options were used but the 'minimum score to output' which was increased to 50 (-T option), resulting in one alignment file of *BAM* format per sample (Figure 2.1). *SAMtools/BCFtools (SAMTOOLS)* (Li *et al.* 2009) version 0.1.18 was used to call SNPs and small indels using default options but the minimum read depth (set to 10) and the maximum read depth (set to 2000). *GATK* (McKenna *et al.* 2010) version 2.8-1 was also used to call both SNPs and small indels using the Unified Genotyper

mode and ploidy equal 1. The overlapping set of variants from the resulting *VCF* files were retained for further analysis. The *GEM* mappability program version 1.315 (Lee & Schatz 2012) was used to calculate mappability values along the whole reference genome using a *k*-mer length of 50bp and 0.04% of allowed substitutions while mapping. Non-unique SNP sites (mappability values greater than one) were filtered out.

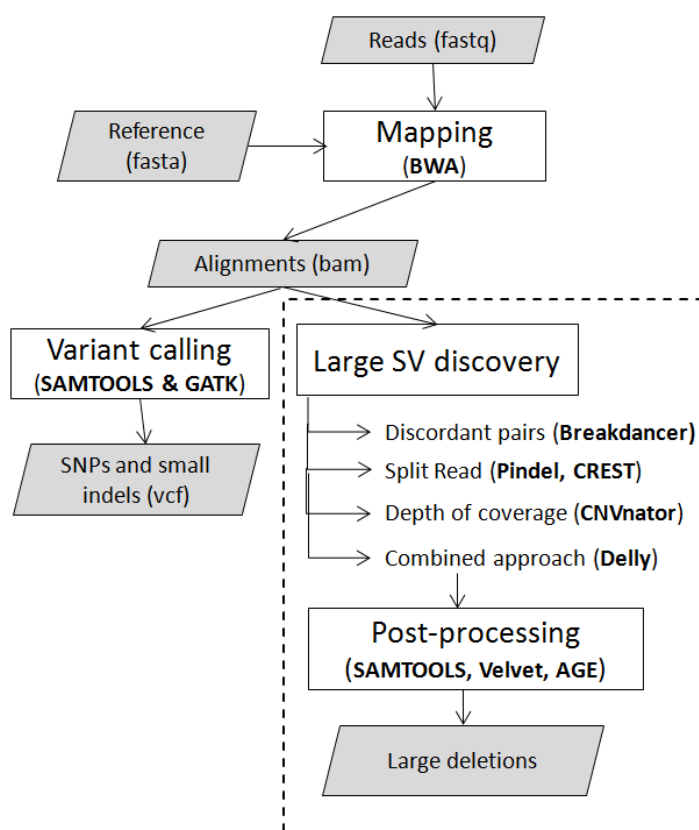


Figure 2.1 Genome variation discovery pipeline

Schematic of the genomic discovery pipeline implemented to compile a catalogue of SNPs, small indels and large deletions from Illumina paired-end sequence data.

Alleles were additionally called at SNP sites using a coverage-based approach. A missing call was assigned if the total coverage at a site did not reach a minimum of 20

or none of the four nucleotides accounted for at least 80% of the total coverage. The sorted proportion of missing calls was plotted for all isolates and a clear inflection point found at around 15%, which was then used as a quality threshold to filter samples (Supplementary Figure 1).

Large deletions (>100bp) were determined using a combination of tools based on paired-end, split-read and DOC approaches. In particular, *Breakdancer* (Chen *et al.* 2009), *CREST* (Wang *et al.* 2011), *Pindel* (Ye *et al.* 2009), *Delly* (Rausch *et al.* 2012) and *CNVnator* (Abyzov *et al.* 2011) were employed followed by a *de novo* assembly-validating strategy. Reads at putative deletions (-/+ 300bp) predicted by all five tools were extracted from alignment (*BAM*) files and subsequently *de novo* assembled using *Velvet* (Zerbino & Birney 2008). If a derived contig happened to be split into two parts when mapping it back to the reference (Abyzov & Gerstein 2011; Camacho *et al.* 2009) with high similarity (>95%), the contig was considered a cross-junction contig (CJC) (Wang *et al.* 2011). Deletions without at least one CJC were considered to be false positives and were therefore discarded. Deletions in PE/PPE genes were filtered out due to the complexity of such regions. These genes are an important source of false positives (Roetzer *et al.* 2013).

All validated deletions were merged when having a mutual overlap greater than 95%. Also, only validated deletion sites predicted by at least two tools or occurring in at least two isolates were retained. The bioinformatic pipeline is summarised in Figure 2.1.

2.2.2 Genomic variation discovery pipeline from complete genomes

A set of 16 publicly available complete *Mtb* genomes were downloaded (Supplementary Table 1). All genomes were aligned against the H37Rv reference genome (NC_000962.3) using *BWA MEM* (Li & Durbin 2010). Once again, SNPs and small indels were identified using *SAMTOOLS* and *GATK*, and the overlapping set of variants retained. Large deletions in complete genomes were derived with an implemented pipeline consisting of *nucmer*, *show-diff* (Kurtz *et al.* 2004) and *AGE* software (Abyzov & Gerstein 2011).

2.2.3 Population structure

Strain spoligotypes for all isolates were derived from *FASTQ* files using *SpolPred* (Section 3.2.2). The best-scoring maximum likelihood phylogenetic tree was computed with *RAxML* v7.4.2 (Stamatakis *et al.* 2008) using all 74,039 SNP sites spanning the whole genome.

2.2.4 PolyTB software architecture

PolyTB has been built using primarily a combination of PHP, HTML and JavaScript code. These three core technologies have become a popular set of tools to develop dynamic web pages because they all are free, open-source and easily combined. PHP is the scripting language working on the server side, normally enclosed inside HTML pages and passed to the PHP parser on the server, which automatically processes it. PHP output is always HTML eventually sent to and displayed by the web browser. JavaScript, on the other hand, is used as the client-side (i.e. Web browser) language. It

provides the means by which elements in an HTML document (such as text boxes, buttons, lists, etc) can be accessed, monitored and changed on-the-fly. JavaScript can gather information from the client and pass it to a PHP script on the server to interactively generate dynamic pages.

Raw data has been initially pre-processed to be stored in the form of PHP and CSV files. Raw files contain information about the samples and populations, genomic polymorphisms (VCF files) and genome annotation. Although PHP has methods to open and read from existing files, having the data already as PHP variables decreases the time required to retrieve it from the server upon request. Such PHP files, hereafter referred to as *data PHP pages*, have no functionality attached to them, and only contain variable assignments. They are located in a server's directory and will be included into, and their variables content used by, another group of PHP pages, the ones actually performing actions on the server, here named *functional PHP pages*. On the other hand, CSV files include data accessible by JavaScript code. The way both groups of PHP pages communicate and the role of JavaScript will be further discussed in the following paragraphs.

The genome browser is the main view of the project; where genetic polymorphisms are displayed for the chromosome region and samples selected by the user. Figure 2.2 shows the web-page structure. JavaScript plays a key role in controlling HTML elements content and layout. The JavaScript code implemented in *browser.js* file contains specific functions to load CSV documents content into *browser.php* HTML elements. For instance, sample names and locations are added as drop-down list

options in this page. JavaScript also monitors changes made by the user on HTML elements (options selected, text written, ticked boxes, etc) and sends the collected information (samples selected, chromosome range, labels ticked, etc) as parameters to *browserparameters.php*. This PHP page acts as a central node by accessing all data (in the form of *data PHP files*, shown in orange and green in Figure 2.2) and keeping the subset meeting the parameters restrictions. For instance, only genes falling within the chosen chromosome range are kept.

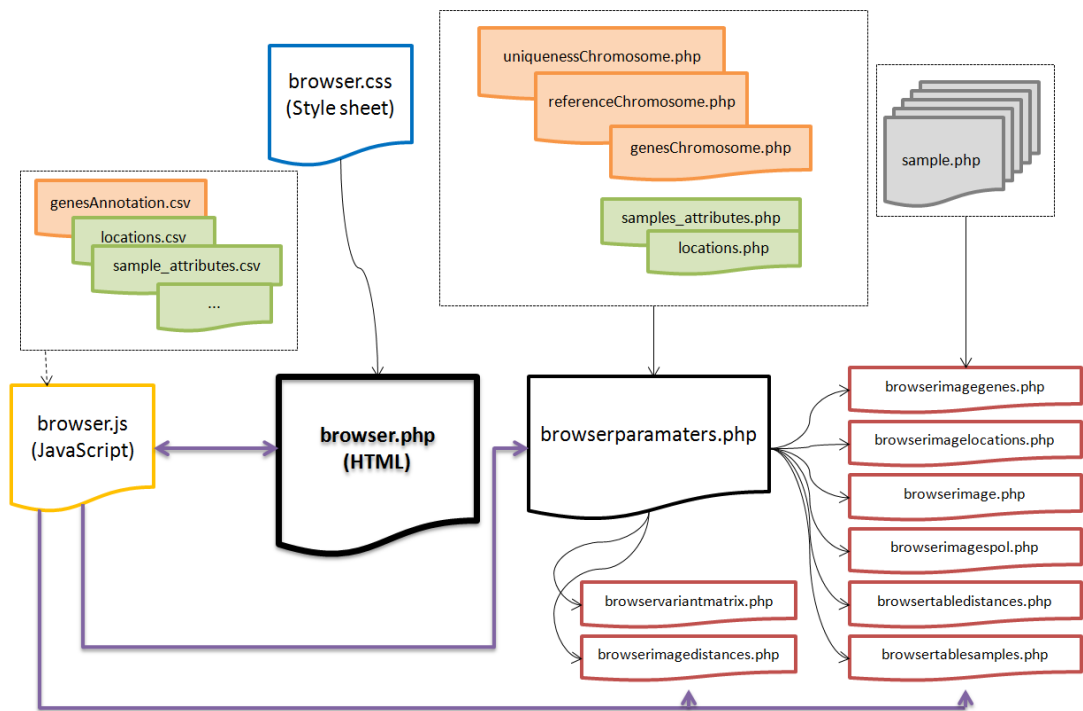


Figure 2.2 Genome Browser View web architecture

Flowchart representing the Genome Browser View architecture. Documents with background colour (PHP and CSV) contain processed data: related to sample metadata (green background), chromosome annotation (orange background) and genetic polymorphisms (*sample.php* files in gray background). This data is accessed by PHP files implementing most of the functionality (white background) including image and table generation. The *browser.js* page implements the JavaScript code which controls the HTML elements in *browser.php*.

The PHP code in *functional PHP pages* (shown in red borders) is executed upon request by JavaScript and their output, HTML tables and PHP images, append it to *browser.php* page. For instance, *browserimage.php* creates an image with colour-coded genetic variants.

The Map View shows allelic frequencies for the chosen variant at the geographical regions from where sequenced samples were collected, either alone or combined with spoligotype frequencies.

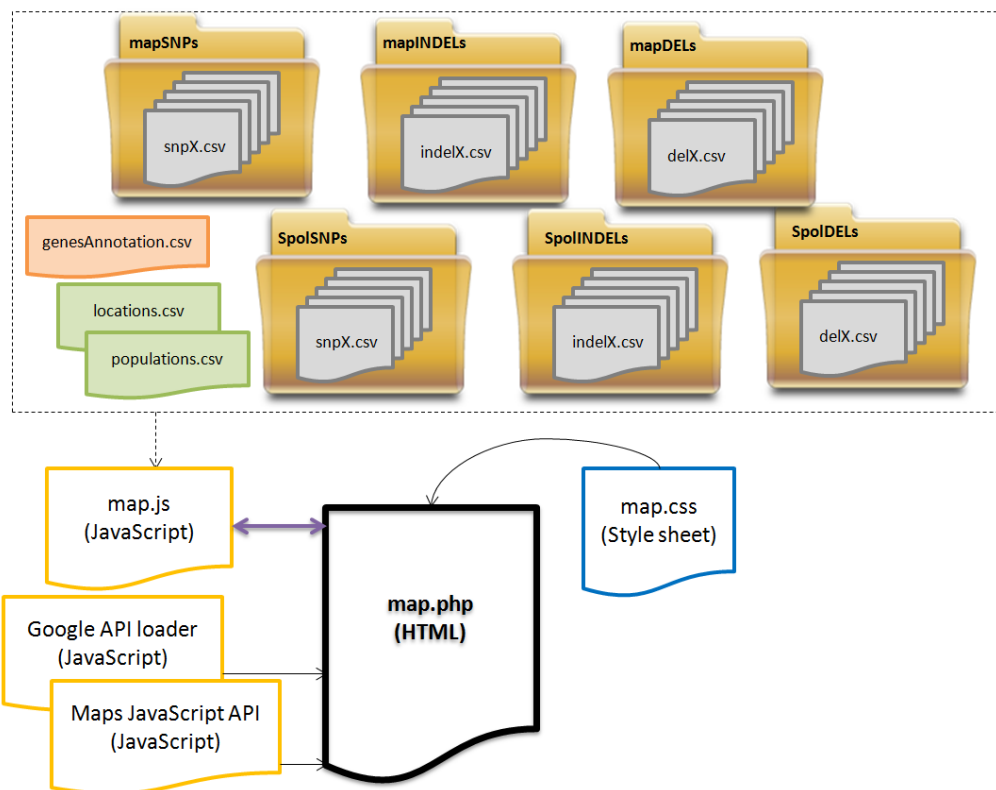


Figure 2.3 Map View web architecture

Flowchart representing the Map View architecture. Documents with background colour (CSV) contain processed data: related to sample metadata (green background) and chromosome annotation (orange background). Allele frequencies across each population are stored in small CSV files (*snpX.csv*, *indelX.csv*, *delX.csv* files in gray background). This data is accessed by JavaScript function in *map.js* which implements most of the functionality and controls the HTML elements in *map.php*.

Figure 2.3 shows the web-page structure. In this case, JavaScript functions in *map.js* implement most of the functionality. Allele frequencies for SNPs, indels and large deletions across geographical populations are stored in CSV files in *mapSNPs*, *mapINDELS* and *mapDELS* folders respectively. For a given selected variant, allele frequencies at all populations are read from its corresponding file and drawn as pie charts on the map. Additionally, spoligotype frequencies within each allele portion can also be displayed in the form of concentric pie charts.

The construction of phylogenetic trees based on whole genome polymorphisms has been implemented as an additional tool to investigate the population structure. The aim is to build phylogenies for the samples selected by the user on the fly and draw the resulting trees in a reasonable amount of time, namely in seconds. Due to the large number of samples and polymorphisms to consider, distance-matrix methods have been chosen for that purpose because of their efficiency. Other methods based on parsimony or maximum likelihood would be more time-consuming when dealing with whole-genome data. However, prior to running the distance method the distance matrix needs to be calculated from multiple alignments, the most computationally expensive step.

As shown in Figure 2.4, the Phylogenetic View keeps the same structure as the Browser View (Figure 2.2). JavaScript functions in *phy.js* collect the samples and parameters specified by the user required for building both the distance matrix and phylogenetic tree and passed them to *createtree.php*. The phylogenetic analysis page required the integration of compiled executables on the server-side and specific

JavaScript libraries. A genetic distance matrix was pre-computed using *PHYLIP dnadist* program from all SNP sites (Felsenstein 1989). Trees are computed on the server upon request by distance-based programs from the PHYLIP package and then displayed on the browser making use of *jsPhyloSVG* JavaScript library (Smits & Ouverney 2010).

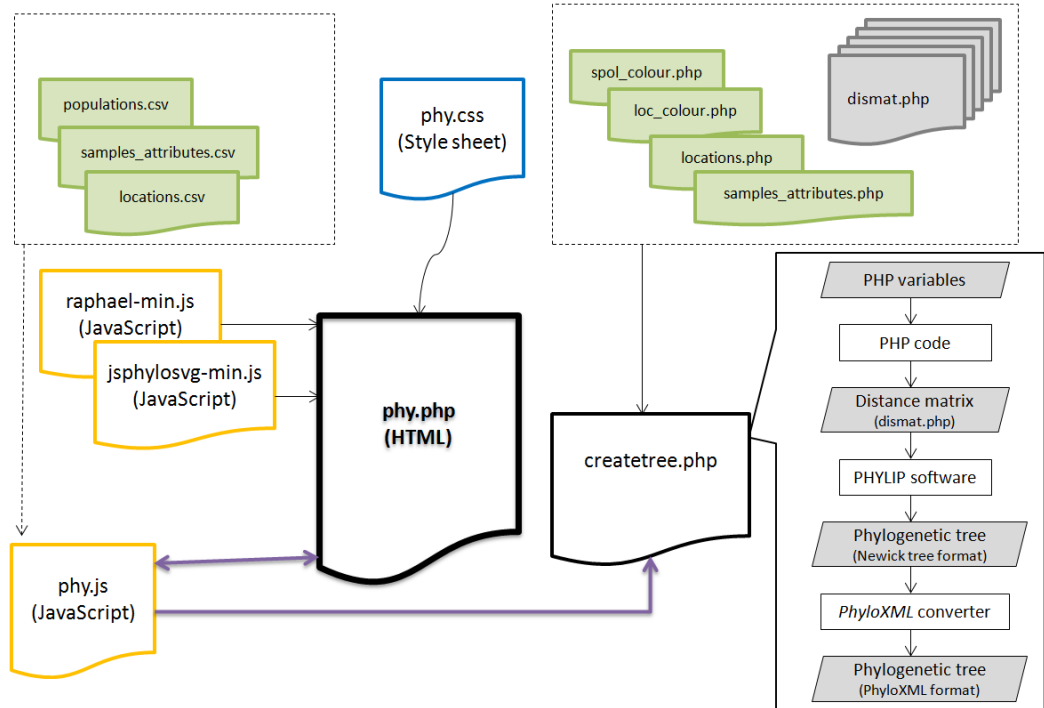


Figure 2.4 Phylogenetic View web architecture

2.3 RESULTS

A high quality SNP dataset ($n = 74,039$) was attained by filtering the list of *SAMTOOLS* and *GATK* consensus variant calls using genomic mappability criteria. Isolates having less than 15% SNP missing calls were retained (1,470/1,627). Both the spoligotypes and lineages were inferred *in silico*, using *SpolPred* software (Coll *et al.* 2012) (Section 3.2.2).

All major modern MTBC lineages are represented, including lineage 1 (East African-Indian (EAI) spoligotype family, 95 isolates, 6.46%), lineage 2 (Beijing, 246 isolates, 16.73%), lineage 3 (Central Asian (CAS), 170 isolates, 11.56%) and lineage 4 (715 isolates, of which 119 X, 273 T, 266 LAM, 7 S and 50 H). Ancestral lineages represented include seventeen *M. africanum* cases, 7 from lineage 5 (West African 1 family), 10 from lineage 6 (West African 2 family) and 6 cases of *M. bovis*. Nearly 15% of isolates (n=218) had orphan spoligotypes, i.e. they were not previously described, but were often closely related to known spoligotypes.

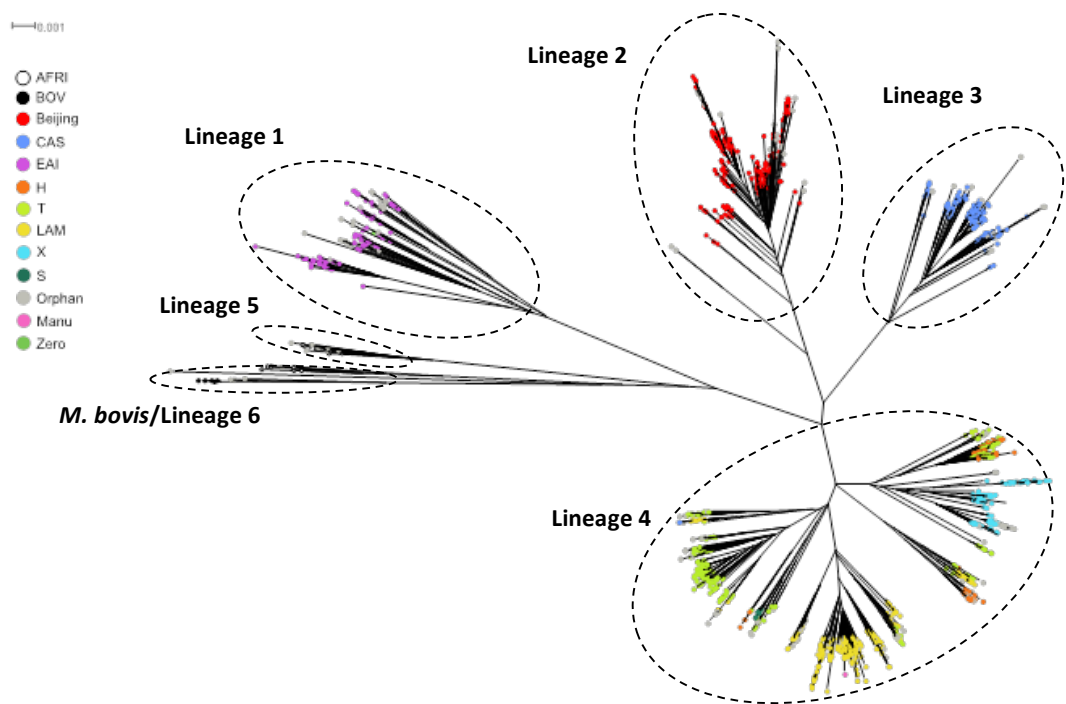


Figure 2.5 *RAxML* phylogenetic tree built for all 1,470 MTBC isolates (colour-coded by spoligotype)

Radial phylogram representation of the best-scoring maximum likelihood phylogenetic tree constructed using *RAxML* software. Samples are colour-coded by spoligotype strain showing a clear correlation of SNP and spoligotype clustering.

Figure 2.5 shows a radial phylogram for all samples, rooted on *M. bovis*. All major MTBC lineages are separated, with *M. bovis*, lineage 1, 2, 3, 5 and 6 isolates clustered within discrete clades, thereby demonstrating the usefulness of SNPs for strain classification. All isolates belonging to lineage 4 are grouped together, although H, T and LAM samples are dispersed among different clades as already observed (Filliol *et al.* 2006).

To highlight the presence of site-specific lineages, edges in the tree were colour-coded by geographical location (Figure 2.6).

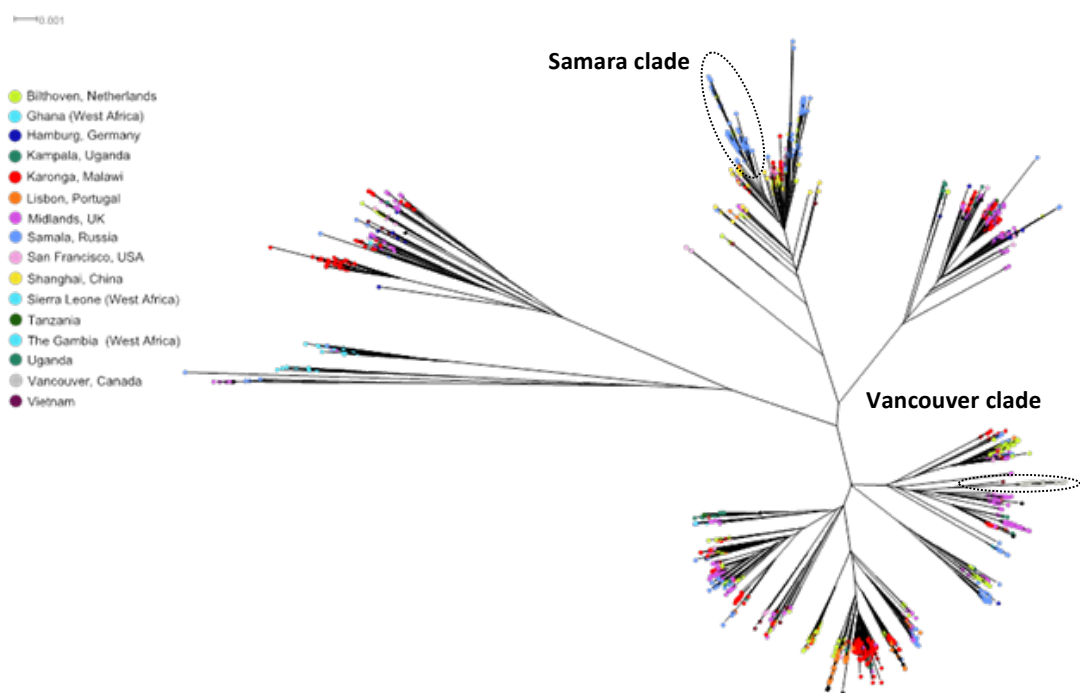


Figure 2.6 *RAxML* phylogenetic tree built for all 1,470 MTBC isolates (colour-coded by geographical location)

Radial phylogram representation of the best-scoring maximum likelihood phylogenetic tree constructed using *RAxML* software. Samples are colour-coded by geographical location to highlight the presence of site specific strains.

The majority of studies (7 out of 8) include isolates belonging to all genetic lineages. In contrast, samples from the Vancouver (SRP002589) study are grouped within the same clade (X spoligotype) suggesting they all resulted from the clonal expansion of the same ancestor (Gardy *et al.* 2011). Similarly, a well-delineated group of Beijing isolates is found to belong exclusively to the ERP000192 study carried out in Samara, Russia (Casali & Nikolayevskyy 2012). The geographical clustering of this sub-group of Beijing isolates corresponds to the “East European” subtype of the Beijing lineage dominant in that region.

2.3.1 Polymorphisms detected and incorporated into *PolyTB*

Of the 74,039 high quality SNPs identified, nearly half (48.9%) were found to be private, namely observed in only one isolate (Figure 2.7).

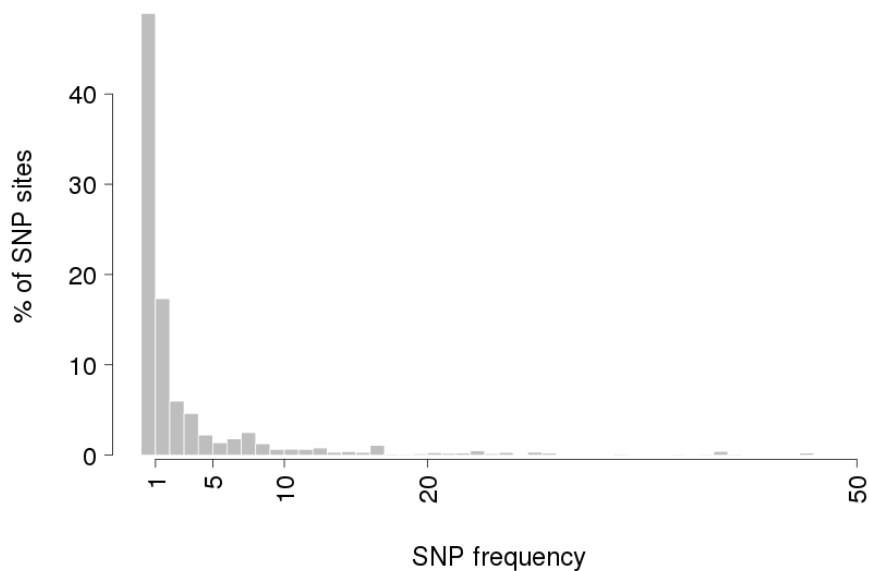


Figure 2.7 SNP frequency bar plot

In general, there were few common SNPs with only 4.6% of SNPs ($n = 3,418$) present in at least 5% of samples. Most of SNPs were found in coding regions of the genome

(median 88.7%) consistent with these regions comprising 91.4% of *Mtb* genome. The majority lead to non-synonymous (NS) changes in amino acids (median 63.0%). Overall, 1,050 SNPs were found per sample on average (range 0 – 2,261 SNPs), corresponding to a median SNP density of 1 SNP per 4.9 kb. SNP density in coding genes (median 0.20, range 0 – 0.50 SNPs/kb) was found to be lower than that in intergenic regions (median 0.27, range 0 – 0.81 SNPs/kb).

Figure 2.8 shows the SNP density calculated across all gene functional categories as annotated in *Tuberculist* (<http://tuberculist.epfl.ch/>). As expected, the highly polymorphic PE/PPE gene families have more SNP density than the average coding regions.

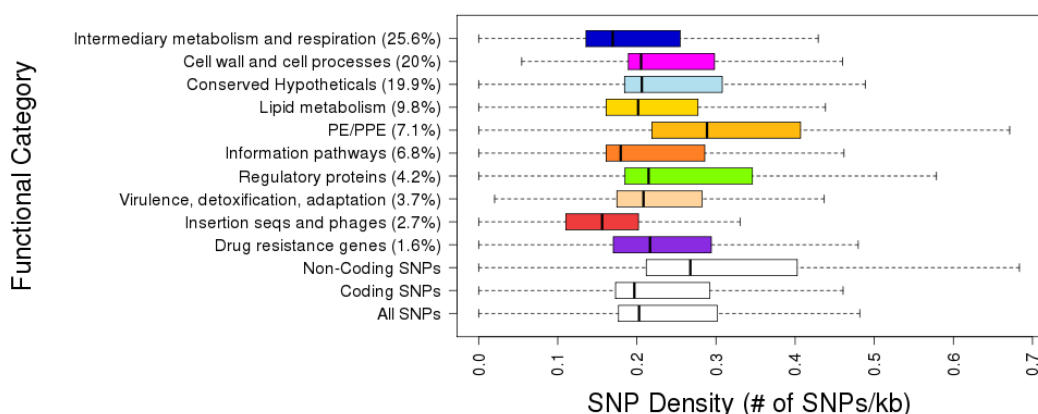


Figure 2.8 Box-and-Whisker SNP density plots by gene functional categories

A total of 4,820 indel loci of size ranging between 1 and 40 bp were identified, with the majority found in single isolates (47.5%) (Figure 2.9). An average number of 85 small indels were detected per sample (range 0 – 199 indels). Both insertions and deletions accounted for an approximately equal proportion of events, 48.8 and 51.2% respectively.

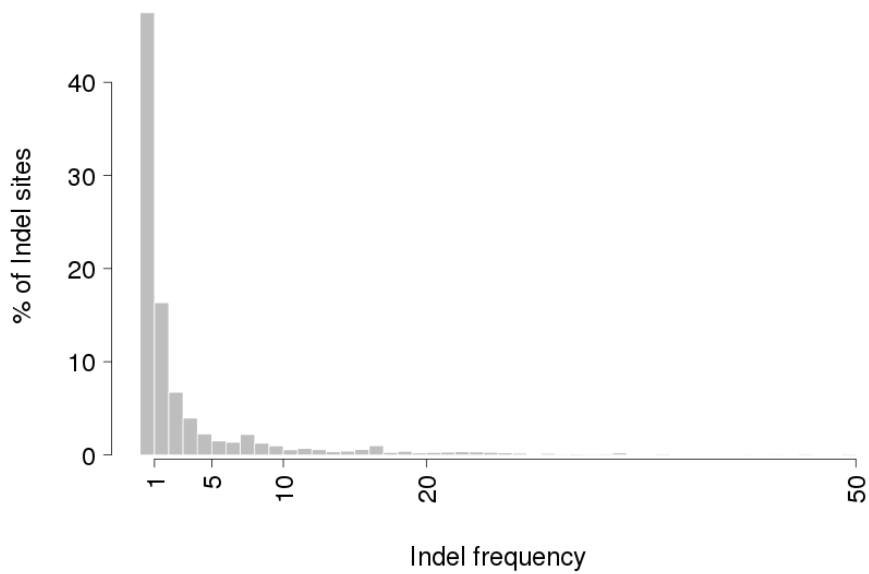


Figure 2.9 Indel frequency bar plot

Indel density was found to be five times smaller in coding genes (median of 1 indel per 83.2 kb) than in non-coding regions (median of 1 indel per 15.7 kb). As was the case with SNPs, the PE/PPE gene families have on average greater indel density than across the rest of coding regions (Figure 2.10).

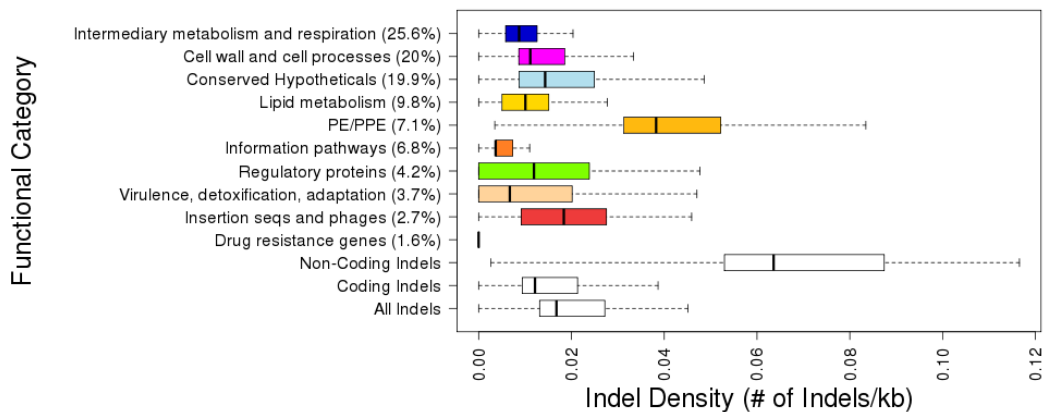


Figure 2.10 Box-and-Whisker indel density plots by gene functional categories

A total of 800 large deletion loci (median size of 541 bp, range 100 – 28,862 bp) were identified after applying a combination of SV detection approaches (pair-end, split-read and DOC) followed by *de novo* assembly and re-alignment validation process of

candidate regions. The average number of deletions per isolate was 18 (range 0 - 38) corresponding to a median density of 1 deletion per 232 kb. Deletion density in coding regions, including those covering whole genes or partially, was 17 times smaller (1 deletion per 576 kb) than that calculated for non-coding regions (1 deletion per 34.3 kb).

The validity of polymorphisms was evaluated by considering known variants extracted from a set of publicly available *Mtb* whole genome sequences (Supplementary Table 1). A total number of 12,887 SNPs, 6,749 small indel and 95 large deletion loci were identified from whole genome comparisons of 16 complete *Mtb* genomes against the H37Rv reference. The WGS-derived polymorphisms were compared against this validated dataset finding an overlap of 4,814 SNP, 319 indel and 26 deletion loci, namely WGS-derived variant loci present in at least one of the *Mtb* complete genomes too. These overlapping polymorphisms were found to be more frequent (17.2%, 18.4% and 43.4% of samples for SNPs, indels and deletions) than those not shared with complete genomes (0.3%, 0.6% and 1.0%). Overall, these results indicate set of WGS-extracted polymorphisms encompass the known variants at the high stringency imposed in the calling procedure.

2.3.2 PolyTB and its applications

PolyTB is a web-based resource (<http://pathogenseq.lshtm.ac.uk/polytb>) that has been designed to facilitate the exploration of MTBC genetic variation (74,039 SNPs, 4,820 indels and 800 deletion sites) at a genome and global scale. The tool consists of complementary and integrated genome browser, map and phylogenetic views. The

genome browser shows SNPs, small indels and large deletions, colour-coded and displayed at their respective genomic coordinates for the chromosome region and isolates selected by the user.

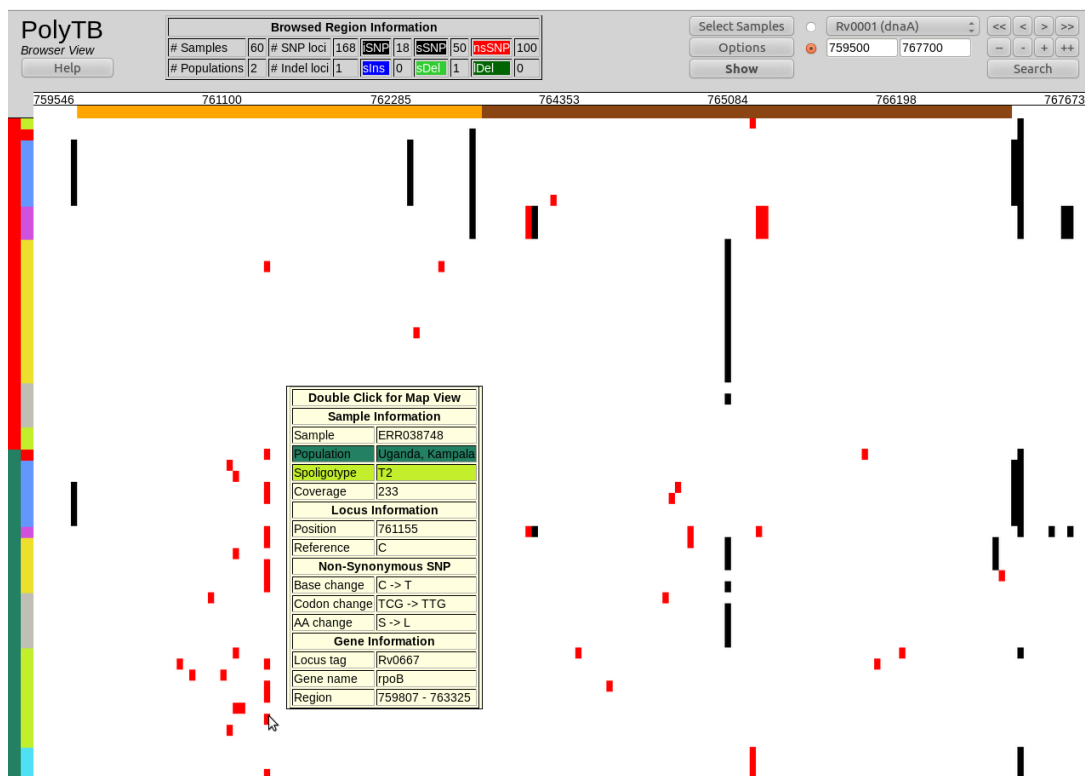


Figure 2.11 Polymorphisms at the *rpoB-rpoC* region (Browser View)

Genetic variants are shown at the *rpoB* and *rpoC* genes, loci known to be associated with RMP resistance. Synonymous SNPs (sSNPs) are coloured in black, non-synonymous SNPs (nsSNPs) in red and small insertions and deletions in blue and green, respectively. Cursor movement over variants displays an information box with further annotation including nucleotide, codon and amino acid changes for SNPs, and length and sequence for indels. *Location* and *Spoligotype* tracks are placed as colour-coded vertical bars at the left hand side of the genomic plot and provide information for samples. Sixty isolates are shown, 30 from Malawi (colour-coded in red in the *Location* bar) and 30 from Uganda (shown in green). Patterns of SNP differences can be observed when comparing isolates from different populations: Kampala isolates harbour many more nsSNPs at *rpoB* gene than Malawian isolates. The observed nsSNPs are likely to be the underlying cause of RMP resistance (Clark *et al.*, 2013). In fact, *rpoB*-516 (A→T SNP at 761,110 bp), *rpoB*-526 (G→T 761,139 bp and A→G 761,140 bp) and *rpoB*-531 (C→G 761,155 bp) mutations are observed in Ugandan isolates, and correspond to nsSNPs already reported as RMP resistance markers (Sandgren *et al.* 2009).

Browsing options allow the user to navigate to the genes or regions of interest, with annotation tracks (top) and sample descriptions (left side) providing context for the variation. *Search functionality* has been implemented to enable the investigation of

polymorphisms at genes of interest given their locus tag, functional annotation, description key words or association with anti-TB DR (Sandgren *et al.* 2009). Figure 2.11 shows differences on polymorphism patterns between isolates from two different populations in the neighbouring *rpoB* and *rpoC* genes, a region associated with RMP resistance. Known RMP resistance markers including *rpoB*-516 (corresponding to the observed 761,110 bp A→T SNP), *rpoB*-526 (761,139 bp G→T and 761,140 bp A→G) and *rpoB*-531 (761,155 bp C→G mutation) are observed in Ugandan isolates. They all correspond to nsSNPs included in diagnostic tests (Bergval *et al.* 2012). Across all populations there are 65 (44 nsSNPs) and 85 (nsSNPs) SNP loci in *rpoB* and *rpoC* genes, respectively.

Users may also consider surveying genomic variants in genes with great importance for the evolution of infection and treatment outcome such as those associated with virulence, nitric oxide production and apoptosis among other possibilities.

Overall, the browser view aims to provide a visualisation tool for the identification of differential variation patterns among isolates and populations at the same region or between different regions under study.

The *map view* shows the global allele distribution for a polymorphism of interest. Allelic frequencies for the chosen polymorphism are displayed as pie charts at the geographical regions from where sequenced samples were collected, either alone or combined with spoligotype frequencies as concentric pies. In the latter, outer arc-sections illustrating strain types are placed on the top of allele frequencies to visually inform of strain type associations with variants at the geographical region investigated.

Figure 2.12 shows an informative SNP (position 4,411,016) found to be associated with lineage 1 (EAI spoligotype family) across studies (only Tanzanian and Karonga-Malawian populations shown). The main purpose of the *map view* is to provide a tool to assess the spread and frequency of WGS-derived genomic variants at a global scale as well as to enable the identification of population- and strain specific polymorphisms.

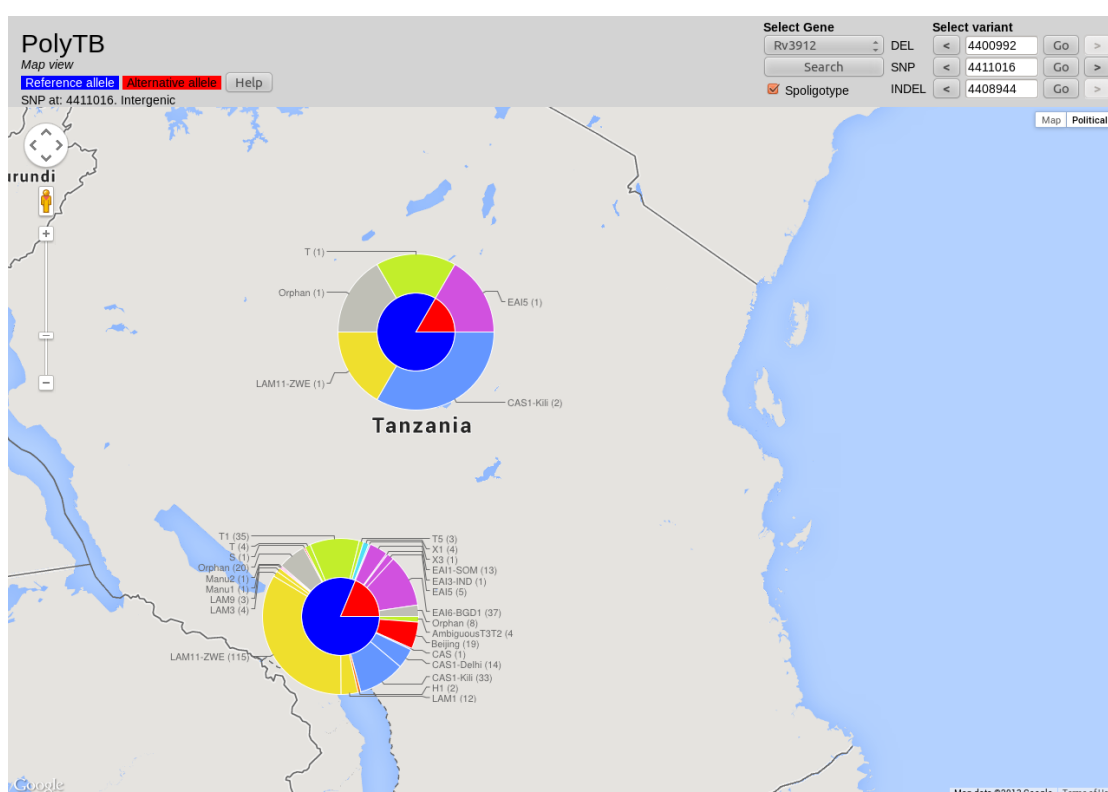


Figure 2.12 SNP associated with lineage 1 (EAI) in Tanzanian and Karonga-Malawian populations (Map view)

Allele frequencies are shown for the chosen polymorphic position as pie charts, either alone or combined with *in silico* inferred spoligotypes (Coll *et al.*, 2012) to allow the visual detection of relationships between certain alleles and strain types. Reference allele frequency portions on pie charts are coloured in blue while alternative allele (i.e. non-reference) frequencies are shown in red. Outer chart portions representing relative strain type frequencies are colour-coded by main spoligotype families (AFRI, BOV, Beijing, CAS, EAI, LAM, Manu, S, T and X). In this particular case, the SNP at 4,411,016 bp position is found to be associated with lineage 1 (EAI) strains in Tanzania and Karonga (Malawi) populations, visualised as the red portion of the inner pie chart linking with the purple portions of the outer pie in both settings.

The *phylogenetic view* allows the user to construct phylogenies for a subset of isolates using whole-genome spanning SNPs. Spoligotypes are included to investigate whether clustering based on SNPs correlates with a strain-type. Figure 2.13 shows the resulting SNP-based neighbour-joining phylogenetic tree constructed for 140 isolates belonging to four different locations.

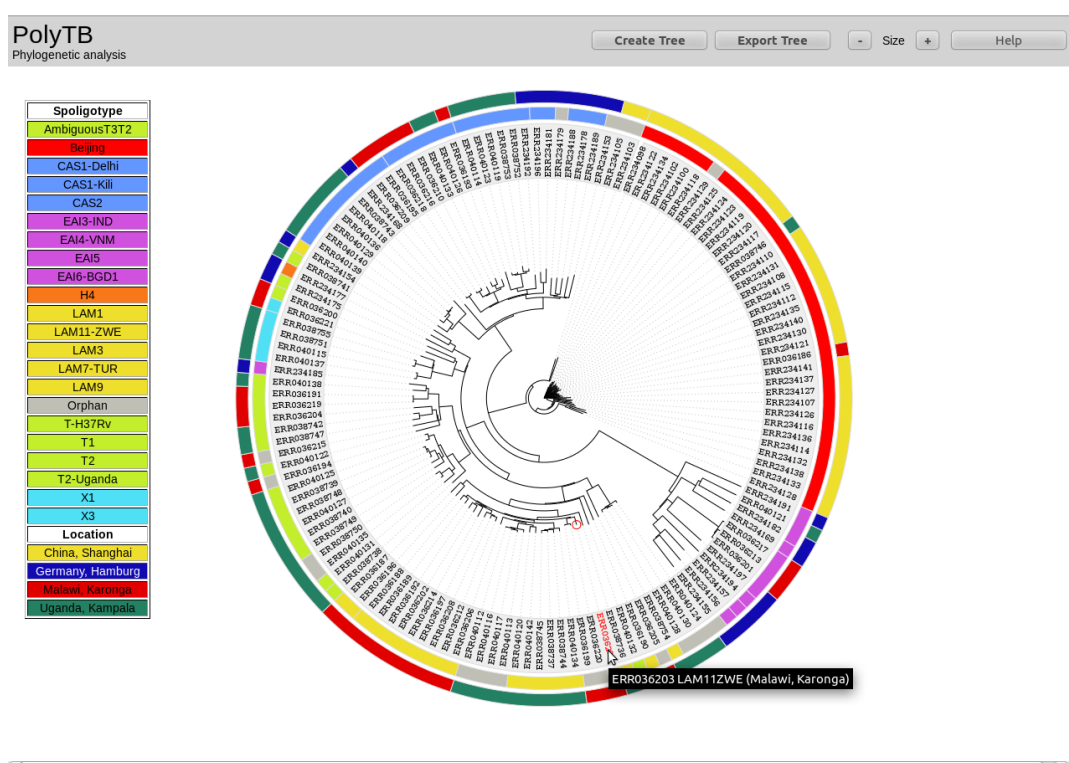


Figure 2.13 SNP-based neighbour-joining phylogenetic tree of 140 isolates belonging to four different locations (Phylogenetic view)

A neighbour-joining phylogenetic tree based on pre-calculated SNP distances is built in real time for the set of 140 isolates from Shanghai (China), Hamburg (Germany), Karonga (Malawi) and Kampala (Uganda). Spoligotype lineages and locations are colour-coded as bar charts around the tree (outer bar representing locations and the inner one spoligotypes) to enable the visual identification of correlations between spoligotype/location and phylogenetic clustering. A table summarising all colour codes will be shown at the left hand side of the page.

Other PHYLIP distance-based methods (Fitch-Margoliash, UPGMA and Least Squares) are available too. Lineages and locations are shown as colour-coded bar charts around the tree to highlight the correlation between lineage and location with phylogenetic clustering. The aim of the phylogenetic view is to assess the genetic relatedness of

isolates within and across populations as well as comparing genetic clustering with spoligotype and geographical assignation.

The *PolyTB* views are linked. For example, the map view is opened for a specific variant position when such a position is double-clicked on the browser view. Likewise, *PolyTB* is linked to external databases. DR genes were extracted from *TBDreamDB* (Sandgren *et al.* 2009), a database describing common mutations associated with DR in *Mtb*. Furthermore, if a particular gene is double-clicked on the top annotation track in the browser view, the user will be forwarded to its *Tuberculist* (<http://tuberculist.epfl.ch/>) entry page containing further annotation information.

2.4 DISCUSSION

Although the TB community has available web-based databases to exploit the existing genotyping data for MTBC (Lew *et al.* 2012), there is no such tool gathering the increasing amount of genetic polymorphisms derived from WGS projects (Stucki & Gagneux 2012). Given the magnitude of the genomic data being generated on a routine basis, efforts must be focused on analysing and presenting this data in a robust and useful manner for the research and public health communities. In this sense, the present release of *PolyTB* makes it the largest open-access repository of genetic polymorphisms derived from WGS projects. The expandable database goes beyond SNPs, and includes small indels and large deletions derived by employing the state-of-the-art variation discovery software. Robust quality control and standardised procedures applied across samples ensures that the datasets are directly comparable. Overall small indel and large deletion densities in coding genes were 5 and 17 times

smaller than in non-coding regions respectively. This considerable reduction in polymorphism density at coding regions can be explained in terms of the potential deleterious effects of these variants in the genome, leading to their selective removal by purifying selection.

Although MTBC strains were historically confined to their endemic geographical locations, migration has led to a more global distribution. Modern modes of transport mean that TB is now easily spread across regions and continents. It is possible to monitor the spread of lineages through phylogenetic markers as well as track DR markers, which emerge *de novo* and independently of strains, with a discriminatory power never achieved before. In this context, the map view provides a tool for the epidemiological surveillance of TB through the geographic distribution of strains and clinically important genetic variants, such as those driving DR. Indeed, knowledge of transmission across lineages and continents is essential to those who need to devise national prevention and control programmes. Similarly, the main purpose of the phylogenetic view is to assess the genetic relatedness of isolates within and across studies as well as comparing genetic clustering with traditional spoligotypes and lineages.

Recently, other tools similar to *PolyTB* have been published. The *tbvar* tool contains 469 isolates and 29,000 SNPs (Joshi *et al.* 2014). Genomic variants (limited to SNPs) are displayed through a table and genome browser views. SNPs are annotated, their functional impact predicted using the SIFT score (Ng & Henikoff 2003) and DR mutations reported. One of the strengths of this tool compared to *PolyTB* is the

'*annoTB*' feature, which enables users to upload their own samples in the form of SNP files. SNPs found in the database are annotated and DR status retrieved based on DR associated mutations in *TBDreaMDB*. The main limitation of *tbvar* compared to *PolyTB* is the lack of sample metadata. Samples lack any lineage or geographical origin information that could help users put the samples into context. Thus, the geographical distribution of SNPs and samples cannot be studied.

Table 2.1 Comparison table of TB WGS genomic databases

Resource (reference)	<i>tbvar</i> (Joshi <i>et al.</i> 2014)	Genome-based <i>Mycobacterium Tuberculosis</i> Variation (GMTV) (Chernyaeva <i>et al.</i> 2014)	<i>PolyTB</i> (Coll <i>et al.</i> 2014)
Number of samples	469	1,084	1,470
Number of studies	37	1	8
Number of polymorphisms	29,000 SNPs	45,655 SNPs and 23,975 indels	74,039 SNPs, 4,820 indels and large 800 deletions
Variants annotated	Yes	Yes	Yes
Views	Tabular and Genome Browser views	Genome Browser and Map views	Genome Browser, Map and Phylogenetic views
Samples metadata	None	DR data, strain type (most of samples) and medical data (minority of samples)	Strain type and geographical data

The *Genome-based Mycobacterium Tuberculosis Variation* (GMTV) database harbours a total of 1,084 MTBC samples and 69,000 variants (SNPs and indels) from Russia. Unlike *tbvar*, this resource contains a broad spectrum of metadata attached to each sample, including TB clinical outcome, year and place of isolation and DR profiles. Genetic polymorphisms can be investigated through a browser and map views, like in *PolyTB*. The main limitation of *GMTV* is that available data is restricted to only one study and therefore only certain lineages are represented. Overall *PolyTB* harbours more samples (1,470) with representatives of all major lineages and more types of

genetic variants (SNPs, indels and large deletions). Furthermore, *PolyTB* includes a phylogenetic view, which enables the study of phylogenetic relationships among samples of the same study and, more importantly, of samples across independent studies. Future extensions of *PolyTB* will incorporate new samples from recently published WGS studies (H. Zhang *et al.* 2013; Casali *et al.* 2014; Pérez-Lago *et al.* 2014; Bryant, Harris, *et al.* 2013) and enhance current functionality, particularly the ‘Search’ feature to allow for complex sample and polymorphism queries.

Current efforts on discovery, visualisation and accessibility of genetic variants from WGS studies must be accompanied with efforts on annotation of these variants (Stucki & Gagneux 2012). This involves identifying strain-specific mutations, which can serve as phylogenetic markers for strain classification, DR-conferring mutations, which are crucial for the development of new and faster diagnostic methods to detect DR, and mutations affecting the bacterial phenotype in various ways, which may have an impact on the outcome of TB infection and disease. In the following chapters these points will be addressed. Strain-specific and DR-associated mutations in MTBC will be identified and their potential use as markers for accurate strain classification and prediction of DR assessed in Chapters 3 and 4 respectively.

RESEARCH PAPER 1

PAPER DETAILS:

Francesc Coll, Kim Mallard, Mark D. Preston, Stephen Bentley, Julian Parkhill, Ruth McNerney, Nigel Martin and Taane G. Clark (2012). SpolPred: rapid and accurate prediction of *Mycobacterium tuberculosis* spoligotypes from short genomic sequences. *Bioinformatics*, 28(22), 2991–3. doi:10.1093/bioinformatics/bts544

AUTHORS CONTRIBUTION:

FC and TGC conceived and designed the study. FC developed and tested SpolPred software. MDP helped optimise the performance and speed of SpolPred software. KM carried out the experimental spoligotyping. TGC, NM and RM jointly supervised the project. FC, RM and TGC drafted and finalised the manuscript with contributions from all other authors. RM provided the samples and SB and JP contributed to the sequencing of these. The final manuscript was read and approved by all authors.

RESEARCH PAPER 3

PAPER DETAILS:

Francesc Coll, Ruth McNerney, José Afonso Guerra-Assunção, Judith R Glynn, João Perdigão, Miguel Viveiros, Isabel Portugal, Arnab Pain, Nigel Martin, Taane G Clark (2014). A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nature Communications*, 5, 4812. doi:10.1038/ncomms5812

AUTHORS CONTRIBUTION:

FC and TGC conceived the project. JAG, JRG, JP, MV, IP, and AP contributed to the construction of data. FC analysed the data. RN, NM and TGC jointly supervised the research. FC, RM, and TGC wrote the paper with contributions from all other authors. The final version of the manuscript was read and approved by all authors.

Chapter 3

Strain typing using whole genome sequences

3 STRAIN TYPING USING WHOLE GENOME SEQUENCES

Despite the growing consensus for the use of SNPs as robust phylogenetic markers, a SNP classification system that can discriminate all known circulating strain types has not been developed. Classical genotyping methods are still broadly employed and results from these are not easily comparable with WGS data. In this regard, this chapter describes the development of approaches for extracting classical genotypes from WGS data and the discovery of strain-specific SNPs that can be used to accurately discriminate circulating MTBC strains.

3.1 INTRODUCTION

Infection with bacteria of the MTBC results in a variety of outcomes including latent infection and/or progression to pulmonary or extra-pulmonary manifestations of disease. Such diversity has been historically attributed to host and environmental factors, and the MTBC was previously considered genetically monomorphic in nature (Lin & Flynn 2010). However, the development of typing methods that discriminate strains into distinct lineages and sub-lineages has demonstrated previously unrecognized diversity. It has been shown that strain type may play a role in disease outcome, variation in vaccine efficacy (López *et al.* 2003) and emergence of DR (Ford *et al.* 2013). Different strains of MTBC have produced distinct biological responses in experimental models and can affect clinical presentation (Nahid *et al.* 2010; Thwaites *et al.* 2008; Caws *et al.* 2008). Strain type may also influence disease epidemiology as in some settings it is associated with the presence or absence of clustering due to recent transmission (Kato-Maeda & Kim 2010). Lineage-specific differences in the

virulence of clinical isolates have been reported across independent experimental systems with modern lineages, such as Beijing and Euro-American Haarlem strains believed to exhibit more virulent phenotypes compared to ancient lineages, such as East African Indian and *M. africanum* strains (Reiling *et al.* 2013). The molecular mechanisms and genetic factors responsible for the described differences in pathogenesis and virulence remain largely unknown. Their investigation requires transparent, easily applied, reliable methods for determining strain type.

The incorporation and standardisation of PCR-based genotyping techniques entailed a turning point in the detection and differentiation of MTBC, allowing the comparison of isolates between laboratories and regions worldwide. Over the last two decades, molecular typing methods such as IS6110-RFLP (Yuen *et al.* 1995), spoligotyping (Kamerbeek *et al.* 1997) and MIRU-VNTR (Supply *et al.* 2001) have been applied and revolutionised epidemiology of TB, by providing insights into the genetic diversity and population structure of MTBC (Schürch & van Soolingen 2012). Six major global MTBC lineages have been defined (1 Indo-Oceanic, 2, East-Asian including Beijing, 3 East-African-Indian, 4 Euro-American, 5 West Africa or *M. africanum* I, 6 West Africa or *M. africanum* II), distinct from a *M. bovis* clade. Lineages 1, 5 and 6 are considered “ancient”, and 2 to 4 “modern”. A novel phylogenetic lineage of MTBC which appears to be intermediate between the ancient and modern has been described recently in Ethiopia and the Horn of Africa (Firdessa *et al.* 2013; Tessema *et al.* 2013), referred to as lineage 7.

Genotyping has been used extensively with epidemiological data to further understanding of TB (Demay *et al.* 2012). For example, at the individual level, cases of recurrence or treatment failure can be explained in terms of reactivation with the same strain, exogenous re-infection or due to polyclonal infection (Ford *et al.* 2012). At a population level, the origins and transmission dynamics of outbreaks can be determined (Walker *et al.* 2013; Gardy *et al.* 2011; Bryant, Schürch, *et al.* 2013), whilst at a global level, TB genotypic lineages have been defined and used to monitor their geographical distribution (Demay *et al.* 2012).

While providing valuable information, standard genotyping methods have several limitations. First, the repetitive nature of genetic polymorphism used by molecular techniques makes them highly prone to convergent evolution (Comas *et al.* 2009), reducing their usefulness as phylogenetic markers. Second, the discriminative power differs between methods, meaning that results from different techniques are not always comparable (Comas *et al.* 2009). Furthermore, isolates with identical DNA fingerprints have been reported to harbour significant genomic diversity (Niemann *et al.* 2009). Therefore standard genotyping tools, which are based on less than 1% of the genome, may not be able to accurately resolve transmission chains and distinguish disease relapse from exogenous re-infection conclusively. On the contrary, SNPs and large sequence polymorphisms (LSP) are ideal markers for defining phylogenetic relationships. The low mutation rate (Schürch *et al.* 2010) and resulting limited sequence diversity in MTBC (coupled with the apparent lack of horizontal gene transfer) make independent mutations at the same site very unlikely. Several studies have already proposed particular sets of SNP (Comas *et al.* 2009; Homolka *et al.* 2012;

Feuerriegel *et al.* 2014; Abadia *et al.* 2010; Stucki *et al.* 2012) and LSP (Gagneux *et al.* 2006) markers to construct reproducible and unambiguous phylogenies in MTBC. Given the predominantly clonal population structure of MTBC, they all produce largely congruent phylogenies (Gagneux & Small 2007). Spoligotype strain classification is also comparable to that assigned by LSP and SNP markers. Indeed, spoligotype families often appear to be sub-lineages within the main six lineages (Kato-Maeda & Gagneux 2011). The effectiveness of the proposed systems is compromised by the limited genetic variation, small numbers of strains or a lack of sub-lineage strain diversity used in their construction.

Given the growing consensus on the use of SNPs as robust and highly discriminatory phylogenetic markers, a new SNP-based classification system is required that can overcome the limitations of current genotyping methods. At the same time, *in silico* genotyping approaches are required to bridge the gap between classical genotyping and high throughput sequencing.

3.2 METHODS

3.2.1 Whole genome datasets and sequence analysis

The raw sequence data of 1,804 MTBC isolates (WGS data set 2) available in the public domain were downloaded from the ENA (<http://www.ebi.ac.uk/ena/>). The analysis of the raw sequence data used is explained in Section 2.2.1. In brief, all isolate sequence data were mapped to the H37Rv reference genome using *BWA* (Langmead *et al.* 2009). *SAMtools/BCFtools* (Li *et al.* 2009) and *GATK* (McKenna *et al.* 2010) were employed to call SNPs and mappability values (Derrien *et al.* 2012) used to filter out non-unique SNP

sites resulting in 91,648 SNP sites. Isolates having less than 15% SNP missing calls were retained (n = 1,601).

3.2.2 *In silico* determination of spoligotypes patterns from short genomic sequences

The popular spoligotyping approach is a genotyping technique that exploits the polymorphism harboured at the direct repeat locus of *Mtb* (Kamerbeek *et al.* 1997). It is based on the PCR amplification of 43 short unique sequences (termed spacers) found between well-conserved 36-bp direct repeats and the subsequent hybridisation of the products onto a membrane with oligonucleotides complementary to each spacer. Since strains vary in the occurrence of particular spacers, each sample produces a distinctive spot pattern then translated into a numerical code of 15 digits, known as octal code (Figure 3.1).

The strategy implemented to derive the spoligotype patterns (i.e. octal codes) from sequence data consisted of screening raw reads and avoided time-consuming post-processing steps like *de novo* assembly to reconstruct the genome sequence. Effectively, even the shortest reads produced by early Illumina sequencing instruments (which were 35-pb long) are expected to span the 25-bp long spacer sequences if they are present in the sequenced genome. In that regard, a C++ program, named *SpolPred*, was developed to predict the spoligotype octal code from files of *FASTQ* format.

By making use of a 2-bit per nucleotide coding strategy to speed up performance, every 25-bp unique spacer is queried against each read allowing up to one mismatch (Ioerger *et al.* 2009). The read length can be changed to support data from different sequencing platforms, such as Sanger-capillary, 454 or AbiSolid. The appearance of all

proportionally with read coverage. Nevertheless, processed reads per unit of time remained constant (approximately 500,000 reads per minute). Once *Spolpred* software accuracy and performance were tested, the tool was run for each of the samples in the global collection (n=1,601) to determine their octal codes and associated spoligotypes.

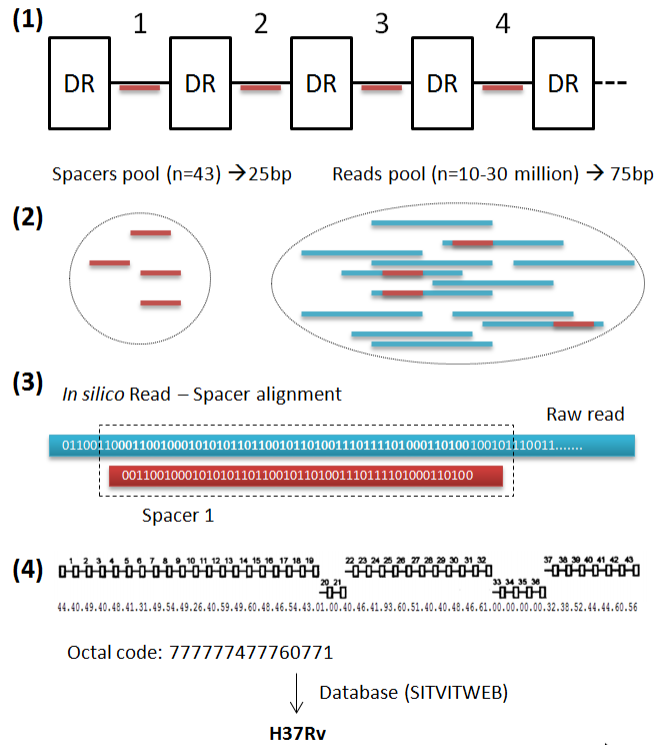


Figure 3.2 *In silico* spoligotyping

(1) Structure of the direct repeat locus in *Mtb* genome. (2) Illustration of a few reads (in blue) containing 25bp spacers (in red). (3) Read-spacer binary alignment performed by *SpolPred*. (4) Resulting number of read-spacer matches and corresponding octal code. Spoligotypes are derived using the latest international spoligotype database (SITVITWEB)(Demay *et al.* 2012).

3.2.3 *In silico* determination of lineages by regions of difference

Lineage-specific LSPs or RDs types were identified from (Gagneux *et al.* 2006) (Table 3.1). This system makes use of 19 phylogenetically informative and lineage-specific deletions to define the global population structure of MTBC, which is often regarded as the gold-standard classification system for MTBC lineages. All samples in the global collection were also genotyped based on this system. Reads covering RDs (+/- 300bp)

were extracted from alignment (*BAM* format) files and subsequently *de novo* assembled using *Velvet* (Zerbino & Birney 2008). If an assembled contig was split into two parts when mapping it back to the reference, with high similarity (>95%) and leaving a gap of length equal to the expected RD length, the contig was considered a cross-junction contig (Wang *et al.* 2011) and the presence of the RD deletion reported.

Table 3.1 Phylogenetically informative deletions

RD	Start	End	Deletion length	Genes involved	Lineage name	Lineage number (1-6)
105	79567	83034	3467	Rv0071-Rv0074	East-Asian	2
115	453364	455971	2607	Rv0376c-Rv0378	Americas-Europe/Euro-American	4
122	669793	670964	1171	Rv0576	Americas-Europe/Euro-American	4
142	1332182	1335033	2851	Rv1189-Rv1192	East-Asian	2
150	1896862	1899349	2487	Rv1671-Rv1674c	East-Asian	2
174	2237049	2240699	3650	Rv1992c-Rv1997	West-Africa/Euro-American	4
181	2535429	2536140	711	Rv2262c-Rv2263	East-Asian	2
182	2545194	2551674	6480	Rv2270-Rv2280	Americas-Europe/Euro-American	4
183	2585853	2588770	2917	Rv2313c-Rv2315c	Americas-Europe/Euro-American	4
193	2704306	2704807	501	Rv2406c-Rv2407	Americas-Europe/Euro-American	4
207	3120521	3127920	7399	Rv2814c-Rv2820c	East-Asian	2
219	3448504	3451396	2892	Rv3083-Rv3085	Americas-Europe/Euro-American	4
239	4092077	4092919	842	Rv3651	Indo-Oceanic	1
702	216795	218516	1722	Rv0186	West-African-2	6
711	1501713	1503655	1943	Rv1333-Rv1336	West-African-1	5
724	2265112	2266239	1128	Rv2018-Rv2019	Central-Africa/Euro-American	4
726	3904958	3906706	1749	Rv3485c-Rv3487c	West-Africa/Euro-American	4
750	1710767	1711556	790	Rv1519-Rv1520	East-African-Indian	2
761	1502787	1503881	1094	Rv1334-Rv1336	South-Africa/Euro-American	4

3.2.4 Phylogenetic analysis

The best-scoring maximum likelihood phylogenetic tree was computed using *RAxML* v7.4.2 (Stamatakis *et al.* 2008) based on 91,648 sites spanning the whole genome. Given the considerable size of the dataset (1,601 samples x 91,648 SNP sites), the rapid bootstrapping algorithm (N=100, x=12345) combined with maximum likelihood search was chosen to construct the phylogenetic tree. The resulting tree was rooted on *M. canettii* (Genbank accession number: NC_019950.1) and nodes were annotated. Subsequently, the ancestral sequence at all internal nodes was computed using *DnaPars* from the *Phylip* package (Felsenstein 1989). The main lineage and sub-lineage defining nodes were initially identified by integrating the topology of the SNP-based phylogenetic tree with the presence of particular RDs and spoligotype composition of the clade. Bootstrap values were computed to assess the confidence of each clade and ensure that all lineage-defined nodes were highly supported (95-100%). For comparison, *RAxML* trees were constructed for the 1601 samples from alignments using other proposed lineage-informative SNPs (Filliol45, Comas93 and Homolka71).

3.2.5 Identification of clade-specific SNPs and selection of the minimal informative set

For each lineage and sub-lineage, the dataset was split into two populations: one containing all samples descending from the clade-defining node and the other with remaining samples. The F_{ST} measure (Weir & Hill, 200) was then calculated for each SNP to identify markers with complete between-population allele differentiation ($F_{ST} > 0.99$). Similarly, the ancestral reconstructed sequence for the clade-defining node was

compared to its closest ancestral node, and the SNP differences derived. A high-confidence set of clade-specific SNPs was obtained by selecting those at the clade-defining internal node and having F_{ST} values of >0.99 in between group comparisons. To ensure that clade-specific SNPs were also suitable markers for their use in strain typing assays, the following filtering criteria were applied: (1) only synonymous SNPs were retained as they are generally under lower selection pressure, (2) SNPs at non-coding regions were discarded since indels are usually more frequent. The density of small indels and large deletions is five and seventeen times smaller, respectively, in coding regions of the genome compared to non-coding (Coll *et al.* 2014) (Section 2.3.1), and (3) only essential genes were used (Stucki *et al.* 2012).

The set of DR associated genes was compiled from *TBDreamDB* (www.tbdreamdb.com) and recent studies (H. Zhang *et al.* 2013). The list of known epitopes in H37Rv was extracted from the *Immune Epitope Database* (www.iedb.org). The gene functional categories were extracted from *Tuberculist* (tuberculist.epfl.ch).

3.3 RESULTS

3.3.1 *In silico* prediction accuracy of spoligotype patterns

SpolPred was initially applied to 51 Ugandan *Mtb* isolates for which WGS data and experimentally determined spoligotypes (44/51 isolates) were available. *SpolPred*-inferred octal code patterns (and their SIT numbers) matched the experimental ones for 39/44 samples (88.6%) (Supplementary Table 2).

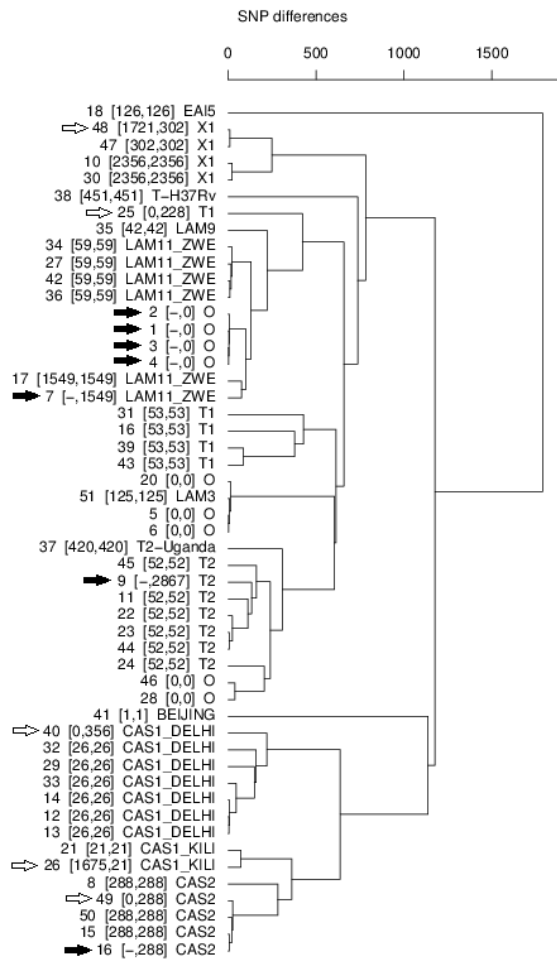


Figure 3.3 Dendrogram for 51 Ugandan isolates constructed using 7k SNPs

Showing from left to right: isolate number, experimentally determined and *SpoIPred* inferred SIT numbers (in square brackets) and *SpoIPred* predicted spoligotype. Isolates with no laboratory data are pointed by black arrows and unknown SIT numbers indicated by dash symbols. Isolates with no matching experimental and *in silico* SIT number are pointed by white arrows.

The five non-matched *in silico* and experiment results were due to the increased *in silico* sensitivity of the detection of spacer 15 in the five samples and, additionally, spacer 26 in one sample. When the original hybridization blots were checked, an irregular signal distribution for spacer 15 across all samples was noted. Some signals were either too faint or just not detectable to be manually assigned as being present. Although predicted spoligotypes remained unchanged for samples 26 and 48, the other three (25, 40 and 49), which had octal codes not previously reported in the

SITVITWEB database, were re-assigned to different spoligotypes. These three isolates were consistently clustered in the SNP-based dendrogram, i.e. within a clade of samples having the same experimental type. Similarly, all samples with no laboratory data were clustered with isolates of the same predicted spoligotype (Figure 3.3). These results demonstrate that *SpolPred* can be employed to accurately and quickly confirm experimentally determined spoligotypes, infer them from sequenced isolates with no laboratory data and reveal unexpected cases of wrongly assigned types.

3.3.2 Population structure of the global collection of MTBC strains

Genomic analysis was performed on whole-genome sequences of 1,601 MTBC isolates from eleven independent sequencing studies from different areas of the world (WGS data set 2), with representation of all seven major lineages (1, n=121, 7.6%, 2, n=390, 24.3%, 3, n=189, 11.8%, 4, n=856, 53.5%, 5, n=17, 1.1%, 6, n=11, 7, n=6), as well as *M. bovis* (n=11) (Supplementary Table 3). A total of 91,648 SNPs were identified, 54.6% were observed in a single sample (Supplementary Figure 2), 89.2% were in coding regions, and 63.5% resulted in non-synonymous changes in amino acids. The SNP-based phylogenetic tree demonstrated a clustering largely congruent with published MTBC phylogenies (Figure 3.4). MTBC main lineages (1-7 and *M. bovis*) and sub-lineages were subsequently identified based on the spoligotype and RD composition of the clades in the SNP-based phylogeny (see Section 3.2.4). The phylogeny revealed the presence of new clades for which RDs do not discriminate. In particular there were gaps in the Euro-American lineage for which molecular fingerprint classifications are less accurate (Comas *et al.* 2009).

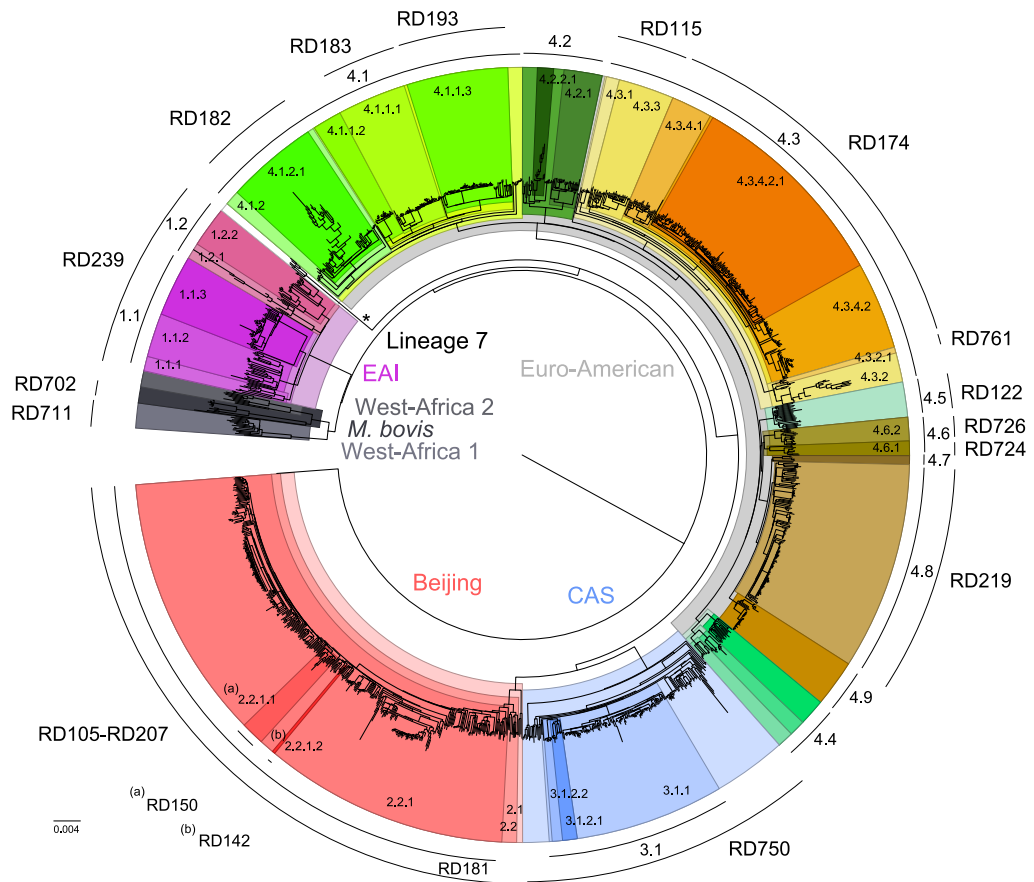


Figure 3.4 Global phylogeny of 1,601 MTBC isolates

A total of 91,648 SNPs spanning the whole genome were used to reconstruct the phylogeny of 1,601 MTBC isolates. All seven main MTBC lineages are indicated at the inner area of the tree. The main sub-lineages are annotated at the outer arc along with lineage-specific regions of difference (RDs). Identified clades are colour-coded.

Although estimates of genetic diversity may be influenced by sampling bias, the greatest nucleotide diversity was observed in lineages 1 (nucleotide diversity $\pi = 0.0103$) and 6 ($\pi = 0.0093$), and the least within lineage 2 ($\pi = 0.0039$) and 3 ($\pi = 0.0040$) strains (Table 3.2). Although spoligotypes tended to cluster within specific clades, there was some evidence of homoplasy, particularly in lineage 4 (Supplementary Figure 3). These anomalies arise from convergent evolution of CRISPR-based spoligotyping polymorphisms (Comas *et al.* 2009).

Table 3.2 Lineage characteristics

Lineage	1	2	3	4	5	6	7
<i>n</i>	121	390	189	856	17	11	6
Description	East-African-Indian	East-Asian	Indo-Oceanic	Euro-American	West Africa 1	West Africa 2	Lineage 7
Spoligotypes	EAI	Beijing	CAS	T, H, LAM, X, S	AFRI_2, AFRI_3	AFRI_1	Unknown
Total SNPs	14,661	22,490	9864	38,033	4761	4560	2458
Avg. SNPs/genome (range)	1970 (1573 - 2094)	1322 (1041 - 1403)	1314 (1046 - 1389)	746 (15-962)	1955 (1831-2030)	2064 (1932-2157)	1962 (1921-1997)
Mean SNPs to MRCA	93	69	53	43	185	207	90
Diversity π	0.0102	0.0039	0.0040	0.0077	0.0058	0.0093	0.0035
RDs present	239	105, 207, 181, 150, 142	750	182, 183, 193, 122, 726, 219, 761, 115, 174, 724	711	702	None
No. Sub-lineages	8	6	5	36	0	0	0
No. Informative SNPs	473	106	262	114	372	220	898
Coding SNPs (% NS)	419 (62.3)	91 (69.2)	231 (64.1)	100 (69)	336 (65.2)	196 (58.2)	808 (64.1)

Abbreviations. MRCA: most recent common ancestor; RD: region of difference; NS: non-synonymous.

All previously reported lineage-specific LSPs or RDs (Gagneux *et al.* 2006) were detected, and their distribution was consistent with clades in the SNP-based phylogeny (Figure 3.4). There was no evidence of homoplasy events using LSPs, further demonstrating their robustness as phylogenetic markers.

As expected, isolates from lineage 1 harboured the RD239 deletion and had EAI-like spoligotypes. Two natural sub-lineages designated 1.1 and 1.2 contained distinctive spoligotype compositions (Table 3.3). The Beijing-specific RD105 deletion was restricted to lineage 2, whilst others (RD207, RD181, RD150 and RD142) were observed downstream from the common ancestor (Figure 3.4), defining sub-lineages within

lineage 2.2. Isolates belonging to lineage 3 harboured the CAS-specific RD750 deletion, and included sub-lineages for CAS1-Delhi (33.9%), CAS1-Kili (53.4%), CAS (6.3%) and CAS2 (5.3%) spoligotypes. All non-CAS1-Delhi samples were grouped into the same clade (sub-lineage 3.1), which was sub-divided into two clades harbouring CAS1-Kili and CAS/CAS2 samples respectively. The Euro-American lineage has been the most poorly characterized historically (Filliol *et al.* 2006), and as expected using spoligotypes there was evidence of non-homogeneous sub-lineages (in particular T, H and LAM families) (see Supplementary Figure 3), potentially due to homoplasmy events (Filliol *et al.* 2006). The phylogeny revealed 36 distinctive clades for lineage 4. All 10 Euro-American lineage RDs were consistently located within one of these clades, further sub-division was achieved and clades with unreported RD were identified (e.g. sub-lineages 4.2, 4.4, 4.7 and 4.9) (Figure 3.4). Representatives of Haarlem (sub-lineage 4.1.2.1), Cameroon (4.6.2), LAM (4.3), S-type (4.4.1.1), TUR (4.2.2.1), Uganda (4.6.1), Ural (4.2.1) and X-type (4.1.1) strains were all identified (Table 3.3). Consistent with previous phylogenetic studies, strains belonging to *M. africanum* were split into West-African lineages 1 and 2, where the latter is phylogenetically closer to the *M. bovis* lineage. Members of the recently described phylogenetic lineage 7 were located as expected at an intermediate location between the ancient and modern lineages.

Table 3.3 MTBC lineages and sub-lineages

Lineage or Sub-lineage	Lineage name	n	Main spol.	Num. Countries	RD	No. clade-specific SNPs
1	Indo, Oceanic	121	EAI	18	239	473
1.1	Indo, Oceanic	82	EAI4, EAI5, EAI6, EAI3	13	239	38

1.1.1	Indo, Oceanic	10	EAI4, EAI5	4	239	57
1.1.1.1	Indo, Oceanic	5	EAI4	1	239	138
1.1.2	Indo, Oceanic	30	EAI5, EAI3	9	239	154
1.1.3	Indo, Oceanic	42	EAI6	2	239	66
1.2	Indo, Oceanic	39	EAI1, EAI2	9	239	5
1.2.1	Indo, Oceanic	11	EAI2	5	239	87
1.2.2	Indo, Oceanic	28	EAI1	4	239	95
2	East, Asian	388	Beijing	14	None	106
2.1	East, Asian	4	Orphan and Manu ancestor	3	None	245
2.2	East, Asian	386	Beijing	13	105, 207	123
2.2.1	East, Asian	376	Beijing	13	105, 207, 181	33
2.2.1.1	East, Asian	16	Beijing	4	105, 207, 181, 150	25
2.2.1.2	East, Asian	2	Beijing	2	105, 207, 181, 142	53
2.2.2	East, Asian	10	Beijing	3	105, 207	80
3	East, African, Indian	189	CAS	18	750	262
3.1	East, African, Indian	121	Non-CAS1, Delhi	7	750	1
3.1.1	East, African, Indian	102	CAS1, Kili	5	750	101
3.1.2	East, African, Indian	17	CAS2, CAS	3	750	14
3.1.2.1	East, African, Indian	10	CAS2	3	750	24
3.1.2.2	East, African, Indian	7	CAS	1	750	188
4	Euro, American	856	S, T, X, LAM, H	22	None	114
4.1	Euro, American	226	T, H, X families	14	None	81
4.1.1	Euro, American (X, type)	138	X family	8	None	35
4.1.1.1	Euro, American (X, type)	47	X2	3	183	59
4.1.1.2	Euro, American (X, type)	20	X1	2	None	85
4.1.1.3	Euro, American (X, type)	69	X3, X1	6	193	106
4.1.2	Euro, American	77	T1, H1	9	None	17
4.1.2.1	Euro, American (Haarlem)	63	T1, H1	9	182	87
4.2	Euro, American	54	LAM7, TUR, H3, H4, T1	6	None	165
4.2.1	Euro, American (Ural)	26	H3, H4	2	None	43
4.2.2	Euro, American	28	LAM7, TUR, T1	6	None	34
4.2.2.1	Euro, American (TUR)	11	LAM7, TUR	2	None	97
4.3	Euro, American	304	LAM	11	None	95

	(LAM)					
4.3.1	Euro, American (LAM)	9	LAM9	2	None	92
4.3.2	Euro, American (LAM)	24	LAM3	3	None	102
4.3.2.1	Euro, American (LAM)	4	LAM3	1	761	70
4.3.3	Euro, American (LAM)	38	LAM9, T5	8	115	55
4.3.4	Euro, American (LAM)	232	LAM11, ZWE, LAM9, LAM1, LAM4	9	174	55
4.3.4.1	Euro, American (LAM)	28	LAM1	5	174	30
4.3.4.2	Euro, American (LAM)	204	LAM11, ZWE, LAM9, LAM1, LAM4	7	174	35
4.3.4.2.1	Euro, American (LAM)	142	LAM11, ZWE	5	174	18
4.4	Euro, American	40	S, T1, T2	7	None	54
4.4.1	Euro, American	22	S, T1	5	None	52
4.4.1.1	Euro, American (S, type)	11	S	4	None	81
4.4.1.2	Euro, American	11	T1	3	None	119
4.4.2	Euro, American	18	T1, T2	3	None	114
4.5	Euro, American	24	H3, H4, T1	3	122	143
4.6	Euro, American	26	LAM10, CAM, T2	6	None	16
4.6.1	Euro, American (Uganda)	16	T2, Uganda, T2	3	724	108
4.6.1.1	Euro, American	3	T2, Uganda	1	724	67
4.6.1.2	Euro, American	13	T2	3	724	64
4.6.2	Euro, American	10	LAM10, CAM, T3	3	726	52
4.6.2.1	Euro, American	2	T3	1	726	230
4.6.2.2	Euro, American (Cameroon)	8	LAM10, CAM	3	726	144
4.7	Euro, American (mainly T)	6	T1, T5	2	None	11
4.8	Euro, American (mainly T)	142	T1, T2, T3, T4, T5	11	219	25
4.9	Euro, American (mainly T)	32	T1	4	None	88
5	West, Africa 1	17	AFRI_2, AFRI_3	5	711	372
6	West, Africa 2	11	AFRI_1	4	702	220
<i>M. bovis</i>	<i>M. bovis</i>	11	BOV_2, BOV_1	3	None	47
<i>M.bovis and 6</i>	<i>M. bovis</i> and West, Africa 2	22	BOV_2, BOV_1 and AFRI_1	6	None	167
7	Lineage 7	6		2	None	898

3.3.3 Identification of lineage and sub-lineage specific SNPs and selection of the minimal informative set

From the 91,648 SNPs, 6,915 lineage and sub-lineage informative markers were identified (list available in <http://pathogenseq.lshtm.ac.uk/tbmolecularbarcodedata>). The distribution of functional categories of genes containing the 91,648 and 6,915 SNP sets did not differ (Figure 3.5A). Using the informative SNPs (n=6,915), there was evidence of difference in the distribution of functional categories between lineages, namely a greater proportion of lipid metabolism non-synonymous polymorphism in lineage 2 (Figure 3.5A), consistent with the greater virulence of the Beijing strain (Kato-Maeda *et al.* 2012). Only 88 SNPs were found in DR candidate regions (2 promoters, 21 genes) (Supplementary Table 6). 22 non-synonymous SNPs were found in 16 *M. tuberculosis* antigenic genes with known epitopes (Supplementary Table 7). A disproportionate number are Haarlem specific (25%, *Rv3873/4*), potentially indicative of its high virulence.

Robust SNPs in essential genes with mutations that lead to synonymous amino acid changes were chosen (n=413, 6%), therefore less likely to be under selective pressure. Redundancy of markers was observed for most of the clades, and one representative per group was randomly selected, leading to a minimum set of 62 SNPs for MTBC classification (Supplementary Table 4). Re-construction of a phylogenetic tree using the 62 SNPs for all 1601 samples resulted in a tree with the same number of delineated clades (Supplementary Figure 4).

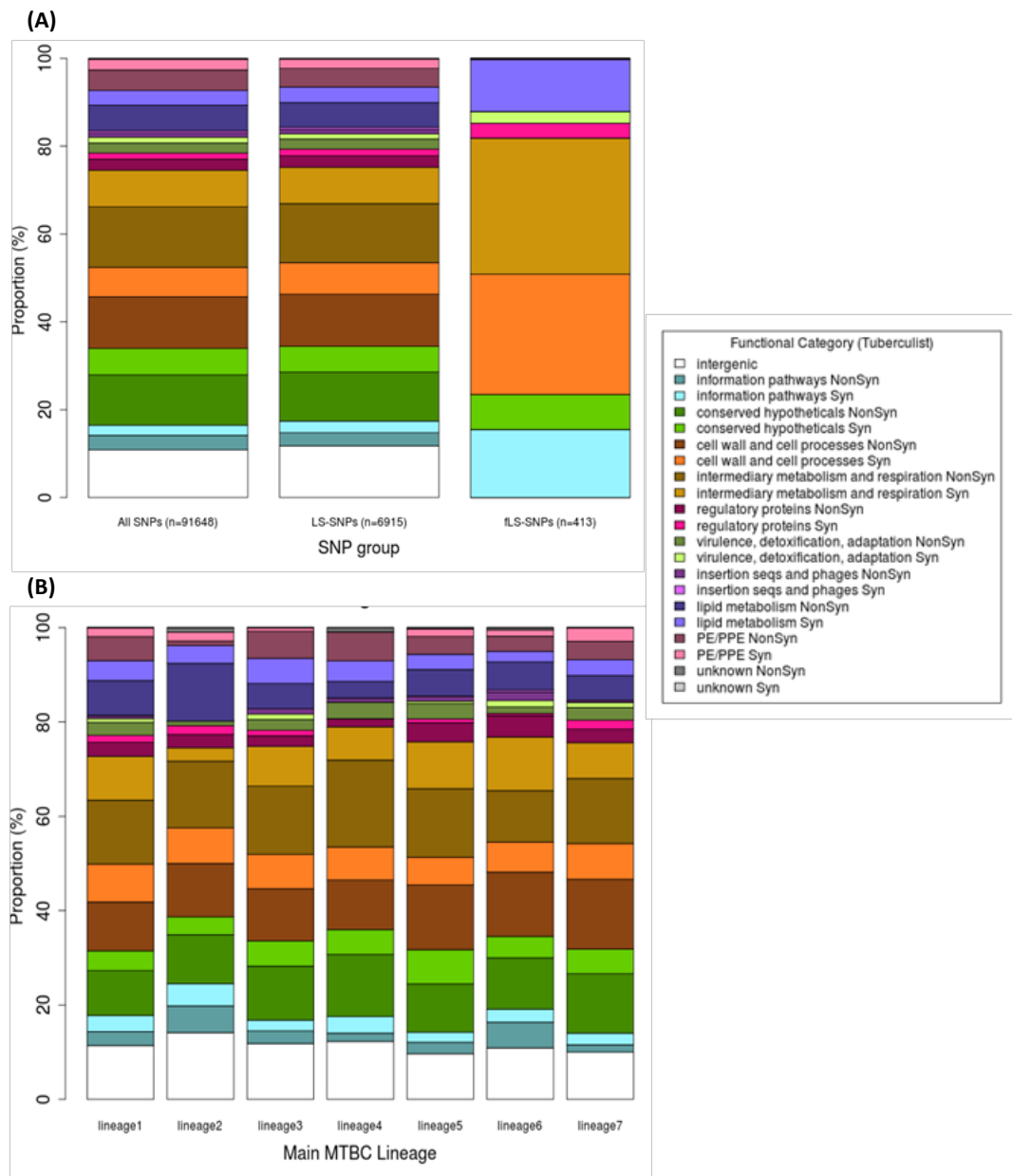


Figure 3.5 Distribution of lineage-specific SNPs across gene functional categories (A) Summary of functional categories for 3 sets of SNPs: all (n=91648, left), lineage, specific (LS-SNPs) (n=6915, middle) and filtered/diagnostic lineage, specific (fLS-SNPs) (n=413, right) (B) LS-SNPs distribution for each of the main six MTBC lineages across gene functional categories (1, 7).

3.3.4 Validation of the proposed SNP-typing system and comparison to other SNP sets

To validate the proposed SNP classification system, it was applied to 27 complete reference genomes representing all MTBC lineages and *M. bovis*, when it was found to predict 100% their reported strain-types (Supplementary Table 5). Furthermore, the scheme was used to classify 850 samples from Samara, Russia, not included in the 1601 samples, and found the same reported lineage proportions (Casali *et al.* 2014). More importantly, unclassified samples from lineage 4 in this study could be assigned to Euro-American sub-lineages by the barcode. A few probable cases of mixed infections were also identified, all combinations of common circulating strain types in that population (Supplementary Table 8).

Lineage-informative SNP sets previously proposed were investigated, denoted here as Filliol45 (45 SNPs (Filliol *et al.* 2006)), Comas93 (93 SNPs (Comas *et al.* 2009)) and Homolka71 (71 SNPs (Homolka *et al.* 2012)). The proportion of these SNPs found among the phylogenetic informative sets differed (Filliol45 29%; Comas93 76%; Homolka71 49%) and some of them were non-segregating across the 1601 samples (Filliol45 17.8%, Comas93 4.3%; Homolka71 39.0%) indicating limitations in their variant and sample ascertainment. Comas93 and Homolka71 sets unambiguously separated the six of the seven main MTBC lineages and the resolved sub-lineages were largely compatible with the ones described in this study (Table 3.4). Still, not all known RD sub-lineages (Gagneux *et al.* 2006), particularly for lineage 4, could be resolved using these classification systems. Phylogenetic trees constructed for the 1601 samples

using these SNP sets (Supplementary Figure 5, Supplementary Figure 6 and Supplementary Figure 7) highlighted the lack of resolution at the sub-lineage level. The proposed set of 62 SNPs are informative for all 7 main MTBC lineages, indicating at least parity in performance with RD typing. Further, the superior number of sub-lineages classified when compared to other SNP systems demonstrates improved strain-type resolution (Table 3.4).

Table 3.4 SNP typing systems comparison

SNP set and reference	Lineage classified (Number of sub-lineages classified)									No. RD lineages covered
	1	2	3	4	5	6	7	<i>M. bovis</i>	Total	
This study 62 SNPs	Yes (8^a)	Yes (6^a)	Yes (5^a)	Yes (36^a)	Yes (0)	Yes (0)	Yes (0)	Yes (0)	7 (55)	19^g
(Homolka <i>et al.</i> 2012) 71 SNPs	Yes (1 ^b)	Yes (0)	Yes (0)	Yes (7 ^d)	Yes (2 ^f)	Yes (0)	No (-)	Yes (0)	6 (10)	NA
(Comas <i>et al.</i> 2009) 93 SNPs	Yes (1 ^b)	Yes (2 ^c)	Yes (0)	Yes (5 ^e)	Yes (0)	Yes (0)	No (-)	Yes (0)	6 (7)	9 ^h
(Filliol <i>et al.</i> 2006) 45 SNPs	No (-)	No (-)	No (-)	No (-)	No (-)	No (-)	No (-)	No (-)	6 (-)	NA

^aSee Table 3.3 for a complete description of lineages and sub-lineages; ^blineage 1.2.1, ^clineages 2.1 and 2.2; ^dlineages 4.6.2.2, 4.1.2.1, 4.3, 4.4.1.1, 4.2.2.1, 4.2.1 and an ambiguous “Ghana”; ^elineages 4.6.2.2, 4.6.1, 4.1.1, 4.1.2.1 and 4.3; ^fWest Africanum Ia & Ib; ^gRD239, 105, 207, 181, 150, 142, 750, 182, 183, 193, 122, 726, 219, 761, 115, 174, 724, 711, 702; 239, ^hRD105, 207, 750, 726, 724, 182, 711, 702 and 7; NA not reported; RD regions of difference

3.4 DISCUSSION

Accurate discrimination between strains of pathogenic bacteria is essential, especially if certain strain types exhibit more virulent phenotypes than others. From an epidemiological point of view, robust, reproducible and highly discriminatory typing systems are required to unambiguously classify clinical isolates, facilitate inter-study comparison and contribute to the control of infectious diseases. Traditional genotyping

methods in TB have been extensively applied in multiple settings, from local outbreak investigations to analysis of the global population structure of MTBC. The introduction of WGS to the study of TB clinical isolates highlighted that significant genomic diversity had been neglected by classical genotyping (Niemann *et al.* 2009). In addition, cases of homoplasy, namely unrelated strains being clustered together due to convergent evolution of their genotyping markers, became apparent. SNPs and other genetic polymorphisms derived from WGS provide enough discriminatory power to unequivocally differentiate strains and are suitable markers for defining phylogenetic relationships.

Although SNPs and other genetic variation derived from sequencing projects are likely to become the genotyping markers of choice, classical genotyping techniques are still widely used and their strain type nomenclature (e.g. Beijing or LAM) broadly employed in the literature. Thus, *in silico* genotyping approaches are required to bridge the gap between experimental and high-throughput sequencing. *SpolPred* achieved high prediction accuracy in the validating dataset; *in silico* derived spoligotypes matched the experimental ones for 39 out of 44 samples. Furthermore, the newly assigned spoligotypes for samples with unknown experimental spoligotype were clustered with other isolates having coincident experimental and *in silico* predicted lineages (Figure 3.3). Interestingly, the absent sequence spacer responsible for the few discrepancies observed, namely spacer 15, was the same across all five problematic isolates. The ambiguous distinction of this spacer has already been reported (Abadia *et al.* 2011) and explained in terms of the presence of a 4-nt deletion adjacent to the amplified sequence (van Embden *et al.* 2000), which would not allow a proper primer

hybridization. Other ambiguities caused by the insertion of IS6110 copies in the direct repeat region have also been reported (Filliol 2000). These results demonstrate that *SpolPred* can be employed to accurately and quickly confirm experimentally determined spoligotypes, infer them from sequenced isolates with no laboratory data (Rashdi & Jadhav 2014) and reveal unexpected cases of wrongly assigned types. Other causes of TB misclassification such as laboratory cross contamination, PCR contamination or ambiguous hybridization patterns could also be clarified.

In order to characterise genome-wide strain specific markers, a genomic analysis was performed on a global collection of 1,601 MTBC isolates. A high-resolution map of polymorphisms consisting of more than ninety thousand SNPs was derived. This genomic variation was used to infer phylogenetic relationships both inter- and intra-lineage to an unprecedented level of resolution, and led to the development of an extendable nomenclature for sub-lineages. All known main MTBC lineages, including the recently discovered lineage 7, and sub-lineages could be identified by integrating spoligotype and RD information with the SNP-based phylogeny. This way the herein described groups can be linked to known RD and/or spoligotype lineages described elsewhere (Table 3.3). An extensive repertoire of 7k lineage and sub-lineage specific SNPs was characterised. The specific genomic variation of known circulating strain-types is likely to contain the genetic factors responsible for lineage-specific phenotypes such as virulence and transmissibility.

A panel of 62 robust SNP markers (of 413 suitable alternatives) was proposed. These markers can be used to construct high-resolution and reproducible phylogenies, be

incorporated in diagnostic assays and assess genotype-phenotype associations. This new genome-wide SNP-typing system entails an advance with respect to previous molecular barcodes for MTBC. These systems are limited due to the small number of genes studied (Homolka *et al.* 2012) or small sample sizes consisting of groups of related strains used in their construction (Filliol *et al.* 2006). The SNP-based phylogenetic tree had higher resolution, and resolved 33 sub-lineages within the historically poorly characterised Euro-American lineage, and added further discrimination within EAI and CAS lineages. There was a high degree of compatibility with this approach and the “gold standard” RD MTBC classification system (Gagneux *et al.* 2006) at the lineage and sub-lineage level (e.g. Haarlem group, lineage 2.1.2.1, RD182; T strains, lineage 4.8, RD115). However, the RD system is incomplete. RD-defined clades (e.g. RD174) harboured multiple SNP-defined groups demonstrating that although phylogenetically robust, LSPs lack resolution. Similarly, some SNP-defined clades, including sub-lineages 4.2 (Ural family) and 4.4 (S-type), lacked a known RD. Spoligotype families (Demay *et al.* 2012) were largely consistent with SNP-based lineages and sub-lineages, but unlike the RD system, cases of homoplasmy events were observed in LAM, T and H strains leading to anomalies (Supplementary Figure 3). The integration of spoligotype and RD data improved the positioning of traditional strain types (e.g. Haarlem, LAM) into SNP-defined sub-lineages within the global phylogeny.

Future work should focus on other types of lineage-specific polymorphisms (e.g. insertions, deletions and large structural variants), which are less common than SNPs, but may have major functional consequences. The proposed system has the flexibility

to incorporate novel strain types should they be reported. The usefulness of the barcode, as an important tool for TB control and elimination activities worldwide, will be enhanced by the incorporation of anti-TB DR mutations, which will be covered in the following chapter (Chapter 4).



Registry

T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

COVER SHEET FOR EACH 'RESEARCH PAPER' INCLUDED IN A RESEARCH THESIS

Please be aware that one cover sheet must be completed for each 'Research Paper' included in a thesis.

1. For a 'research paper' already published

1.1. Where was the work published?

1.2. When was the work published?

1.2.1. If the work was published prior to registration for your research degree, give a brief rationale for its inclusion

.....
.....
.....

1.3. Was the work subject to academic peer review?

1.4. Have you retained the copyright for the work? **Yes / No**

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 copyright holder (publisher or other author) to include work

2. For a 'research paper' prepared for publication but not yet published

2.1. Where is the work intended to be published? Genome Biology

2.2. Please list the paper's authors in the intended authorship order

See next page

2.3. ~~Stage of publication~~ ~~Not yet submitted~~ / Submitted / ~~Undergoing revision from peer reviewers' comments~~ / ~~In press~~

3. 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)

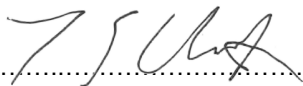
.....
See next page

NAME IN FULL (Block Capitals) FRANCESC COLL I CEREZO

STUDENT ID NO: 323873

CANDIDATE'S SIGNATURE 

Date 10/12/2014

SUPERVISOR/SENIOR AUTHOR'S SIGNATURE (3 above) 

RESEARCH PAPER 4

PAPER DETAILS:

Francesc Coll, Ruth McNerney, Mark D Preston, José Afonso Guerra-Assunção, Andrew Warry, Grant Hill-Cawthorne, Kim Mallard, Mridul Nair, Anabela Miranda, João Perdigão, Miguel Viveiros, Isabel Portugal, Zahra Hasan, Rumina Hasan, Judith R Glynn, Nigel Martin, Arnab Pain, and Taane G Clark (2014). Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Submitted.

AUTHORS CONTRIBUTION:

FC, RM, AP, and TGC conceived and designed the study; FC and MDP developed and tested the online tool; JAG and AW developed additional software; GH-C, MN and KM performed laboratory experiments and curation of meta data for sequencing; AM, JP, MV IP ZH, RH, JRG contributed biological samples, sequencing or phenotypic data; FC performed the statistical analysis under the guidance of NM and TGC; AP led the sequencing efforts; FC, RM and TGC wrote/drafted and finalised the manuscript with contributions from all other authors. The final manuscript was read and approved by all authors.

Chapter 4

A whole-genome sequencing
approach for drug resistance
profiling

4 A WHOLE-GENOME SEQUENCING APPROACH FOR DRUG RESISTANCE PROFILING

The usefulness of WGS for accurate and highly discriminatory strain typing in TB has been repeatedly proved in recent reports (Comas *et al.* 2009; Roetzer *et al.* 2013; Bryant, Harris, *et al.* 2013) and is reinforced by the results presented in the previous chapter (Chapter 3). Although WGS could simultaneously provide other clinically relevant information, the potential use of this technology for detection of DR forms of TB has been largely unexplored. In this regard, this chapter aims to assess the potential of a whole-genome approach to detect DR-TB.

4.1 INTRODUCTION

4.1.1 Drug resistance: a threat to disease control

Resistance has been reported to all drugs used to treat TB (Dheda *et al.* 2014). Increased resistance is associated with decreased patient survival and the emergence of resistance to first and second line drugs is a substantial threat to disease control. The WHO classifies TB resistant to INH and RMP, the two key first-line anti-TB drugs, as multi drug-resistant (MDR-TB), and a switch to second line treatment is advised. Resistance to additional drugs such as EMB, or STR further compromises treatment (Tahaoğlu *et al.* 2001; Migliori *et al.* 2009). MDR strains that have developed resistance to the FLQ and AMI used in second line treatment are classed as extensively drug resistant (XDR-TB). The term total drug resistance (TDR-TB) has been used to describe strains found resistant to all drugs for which tests are available, but there is not yet an agreed definition of TDR-TB (Dheda *et al.* 2014). Treatment of these forms of DR

disease is difficult and expensive, and outcomes are poor (Pooran *et al.* 2013; Abubakar *et al.* 2013; Lange *et al.* 2014). Treatment is long, in some cases years, and involves drugs of heightened toxicity. Adverse reactions are common and may be severe and irreversible (Yee *et al.* 2003). Poor tolerance leads to reduced compliance, which in turn reduces cures rates and can result in amplification of resistance (Shean *et al.* 2013).

4.1.2 Review of mechanisms of drug resistance in *Mycobacterium tuberculosis*

The primary cause of resistance in *Mtb* is the accumulation of point mutations and indels in genes coding for drug-targets or drug-converting enzymes. INH is a pro-drug activated by the catalase-peroxidase enzyme KatG. The active form of this drug blocks the substrate of InhA, an enzyme involved in mycolic acid biosynthesis, essential components of the mycobacterial cell wall, leading to the disruption of the cell wall and resulting in a loss of cellular integrity. Altered or diminished catalase-peroxidase activity caused by nsSNPs in *katG* is the most frequent mechanism of INH resistance (INH^R). Small indels (commonly frameshift mutations) and missense mutations have been observed at relatively low frequency in INH^R clinical isolates and are responsible for high levels of INH^R (Slayden & Barry 2000). Down-regulation of *katG* has also been suggested to be a mechanism of INH^R and mutations in the *furA-katG* intergenic region (the putative *katG* promoter) to be responsible for such resistance (Ando *et al.* 2011). An alternative genetic strategy by *Mtb* to diminish INH interference consists in mutations in the binding site of InhA or over-expression of this enzyme by point mutations in its promoter region (Slayden & Barry 2000). Indeed, the presence of mutations in the *inhA* promoter together with mutations in the *inhA* coding region can

lead to the development of high-level INH^R (Machado *et al.* 2013). Mutations in other loci such as *ahpC* and *kasA* have also been linked to INH^R. The former gene is co-expressed with *katG* in response to oxidative stress. It is not clear though whether an increased expression of *ahpC* is a consequence of the loss of functional *katG*, i.e. compensatory mutations, or a cause of resistance. Similarly, it is unclear whether the KasA enzyme involved in mycolic acid synthesis is also a direct target of activated INH and therefore mutations in its binding site a cause of resistance. On the other hand, the observed *kasA* mutations in INH^R isolates could be in fact compensatory mutations due to the inactivation of the actual target (i.e. InhA) involved in the same metabolic pathway.

In contrast to the multiple reported mechanisms of INH^R, RMP resistance (RMP^R) is determined by mutations in the *rpoB* gene. RMP and other rifamycins have high affinity to the RNA polymerase encoded by *rpoB* and *rpoC* in *Mtb*. While nsSNPs and indels in the coding region of *rpoB*, mostly in the 81-bp RMP resistance-determining region (RRDR), are the main cause of RMP^R, compensatory mutations in the *rpoC* originate to restore the fitness cost caused by *rpoB* mutations (de Vos *et al.* 2013).

The genetic causes of resistance to other first-line drugs are not fully characterised. STR is known to inhibit protein synthesis by interfering in the small 30S subunit of the ribosome, precisely in the 16S ribosomal RNA encoded by *rrs* gene and the S12 ribosomal protein encoded by *rpsL* gene in *Mtb*. Mutations in the *rrs* gene have been linked to intermediate levels of resistance and account for 20% of STR-resistant strains. *rpsL* mutations are generally associated with high levels of STR resistance and are found in 50% of resistance cases (Zhang & Vilcheze 2009). However, strains without

mutations in either of these two genes, frequently presenting low-levels of resistance, have also been reported and thought to harbour mutations in secondary targets such as *gidB* or unknown genes (Moure *et al.* 2013).

PZA, like INH, is a pro-drug which has to be catalysed to be activated, in this case by the pyrazinamidase encoded by *pncA* gene (Scorpio & Zhang 1996). Mutations in the *pncA* coding region or its putative promoter are associated with PZA resistance (PZA^R) and both SNPs and indels have been described (Stoffels *et al.* 2012a). However, not all mechanisms of resistance are currently characterised as PZA^R clinical strains lacking *pncA* mutations are extensively reported. Despite recently uncovered PZA potential targets, particularly *rpsA* (Shi *et al.* 2011) and *panD* (S. Zhang *et al.* 2013), the overall mode of action remains elusive (S. Zhang *et al.* 2013).

Resistance to EMB remains poorly understood despite the multiple genes identified to date. Mutations in the *embCAB* operon, which encodes for enzymes involved in the biosynthesis of arabinan components of mycobacterial cell wall, are responsible for EMB resistance (EMB^R) (Telenti *et al.* 1997). However, many clinical strains have mutations in these genes while remaining susceptible to EMB. It is becoming evident that EMB^R develops through mutations in multiple loci, including *embA/B/C* genes and other currently unknown genes, resulting in a range of different levels of resistance (Safi *et al.* 2013).

Ethionamide (ETH) is a second-line anti-TB drug indicated to treat MDR-TB. Like INH and PZA, ETH is a pro-drug that needs to be activated, in this case by the EthA enzyme (Baulard *et al.* 2000), whose expression is negatively regulated by EthR, a transcriptional repressor that interacts directly with the *ethA* promoter region

(Engohang-Ndong *et al.* 2003). ETH is a structural analogue of INH and shares the same molecular target, namely the InhA enzyme involved in the synthesis of mycolic acids (Banerjee *et al.* 1994). Resistance to ETH has been reported to result from mutations in the ETH-converting enzyme EthA, mutations in the coding region of InhA (resulting in cross-resistance to both ETH and INH), mutations in the *inhA* promoter region leading to an over-expression of the target and cross-resistance to INH; and mutations in the EthR transcriptional regulator (Engohang-Ndong *et al.* 2003). Mutations in these four loci account for 80% of ETH-resistant cases and therefore other mechanisms of ETH resistance remain to be discovered (Brossier *et al.* 2011).

FLQ are currently used to treat TB when resistance to first-line drugs has developed. This family of drugs kill *Mtb* by binding to and interfering with the DNA gyrase, which consists of two sub-units encoded by *gyrA* and *gyrB* genes (Takiff *et al.* 1994). Resistance to FLQ arises from mutations in the quinolone resistance-determining region (QRDR) located within *gyrA* and *gyrB*. In most of studies, more than 90% of FLQ-resistant strains have mutations in the QRDR (Maruri *et al.* 2012). QRDR mutations confer cross-resistance within the FLQ, albeit not at the same level. For the same mutations, moxifloxacin (MOX) normally presents the lowest minimum inhibitory concentration (MIC) values in the group followed by levofloxacin (LEVO), and in contrast with the higher levels of resistance observed for ofloxacin (OFX) and ciprofloxacin (CIP) (Malik *et al.* 2012). These differences explain the better clinical efficacy of MOX (Feasey *et al.* 2011) and LEVO compared to CIP and OFX (Angeby *et al.* 2010).

AMI drugs kanamycin (KAN), capreomycin (CAP) and amikacin (AMK) are second-line injectable antibiotics used to treat MDR-TB. AMI are ribosome-binding antibiotics that target the 16S rRNA encoded by *rrs* gene. Mutations in the 1,400-bp region of *rrs* confer cross-resistance to all AMI members albeit not at the same level. Mutations in the *eis* promoter region have been associated with KAN resistance. Additionally, mutations located throughout the whole *tlyA* gene are only associated with CAP resistance.

4.1.3 Available diagnostic tests for drug resistant tuberculosis

Early detection of DR is crucial for access to effective treatment and prevention of onward transmission. Knowledge of the full drug susceptibility profile would enable tailored treatment to improve efficacy and reduce exposure to ineffective toxic drugs. Current testing for resistance to most anti-TB drugs involves isolation and culture of the bacteria followed by exposure to the drug, a process that takes weeks or months and requires high levels of microbiological safety. Rapid molecular assays are now available for some key drugs that test directly from sputum and in 2013 the Xpert MTB/RIF (Cepheid, Sunnyville, USA) was granted US FDA approval for detecting resistance to RMP, conditional on confirmatory testing by a reference laboratory. This easy-to-use semi-automated PCR-based test has also been endorsed by WHO, as have Line Probe Assays (LiPA) for resistance to RMP and INH (GenoType MTBDRplus Assay), where, following amplification of bacterial DNA samples are interrogated with a panel of oligonucleotide probes (Ling *et al.* 2008). LiPA to detect resistance to other drugs, including FLQ and AMI have also been developed (Genotype MTBDRsl Assay) (Ajvani *et al.* 2012), but have yet to be endorsed by WHO. Though undoubtedly useful, both

technologies are limited in the number of loci they examine and they lack capacity to differentiate silent mutations from those that effect drug efficacy (Alonso *et al.* 2011; Jin *et al.* 2013; Aubry *et al.* 2014).

4.1.4 Whole-genome sequencing for the detection of drug resistance

WGS has the potential to overcome such problems and extend rapid testing to the full range of anti-TB drugs. Sequencing technologies are evolving rapidly and benchtop analyzers have been developed capable of sequencing a bacterial genome in a few hours. Costs have been greatly reduced with the introduction of high throughput technology (Sboner *et al.* 2011). NGS currently assists patient management for a number of conditions (Berg *et al.* 2011; Köser *et al.* 2012; Harismendy *et al.* 2013). The relatively small genome of *Mtb* (4.4 Mb) and its inherent stability render it a suitable candidate for genomic analysis but the complexity of data interpretation has, thus far, restricted whole genome analysis to the research laboratory. Recent reports of sequencing *M. tuberculosis* from sputum from suspected XDR-TB patients suggest this will soon change. However, data analysis remains a bottleneck, requiring specialist expertise not readily available in clinical laboratories. To address this issue and progress sequencing towards real time management of patients a rapid, online tool for analyzing raw sequence data and predicting resistance was developed. Accuracy data is presented for eleven anti-tuberculosis drugs from using the tool to interrogate raw whole genome sequence data from clinical isolates, compared to their phenotype obtained by conventional DST. To assess the potential benefits of a whole genome approach a new library of mutations was curated and its performance compared to

those used in three commercial molecular tests, the Xpert MTB/RIF (Cepheid Inc, USA), and the MTBDRplus and MTBDRsl (Hain Life Science, Germany).

4.2 METHODS

4.2.1 Mutation library

Following review of available data a library of mutations predictive of DR was compiled (see list in <http://pathogenseq.lshtm.ac.uk/rapiddrdata>). Drugs included were AMK, CAP, EMB, ETH, INH, KAN, MOX, OFX, PZA, RMP and STR.

Table 4.1 Summary of mutations included in the curated drug resistance mutation library

Drug	Loci	# of variable sites (SNPs, indels)
INH	<i>katG</i>	241 (286, 25)
	<i>katG promoter</i>	3 (3, 0)
	<i>inhA</i>	14 (17, 0)
	<i>inhA promoter</i>	9 (11, 0)
	<i>ahpC</i>	8 (8, 0)
	<i>ahpC promoter</i>	13 (14, 0)
	<i>kasA</i>	8 (11, 0)
RMP	<i>rpoB</i>	89 (135, 19)
	<i>rpoC</i>	8 (8, 0)
EMB	<i>embB</i>	124 (154, 1)
	<i>embA</i>	5 (5, 0)
	<i>embA promoter</i>	3 (3, 0)
	<i>embC</i>	26 (27, 0)
	<i>embR</i>	22 (24, 0)
STR	<i>rrs</i>	21 (25, 0)
	<i>rpsL</i>	14 (19, 0)
PZA	<i>pncA</i>	215 (270, 64)
	<i>pncA promoter</i>	4 (6, 0)
	<i>rpsA</i>	3 (4, 0)
ETH	<i>ethA</i>	33 (29, 5)
	<i>ethR</i>	3 (4, 0)
	<i>inhA promoter</i>	3 (3, 0)
	<i>inhA</i>	4 (5, 0)
FLQ	<i>gyrA</i>	16 (23, 0)
	<i>gyrB</i>	22 (29, 0)
AMK	<i>rrs</i>	8 (9, 0)
CAP	<i>rrs</i>	3 (4, 0)
	<i>tlyA</i>	26 (18, 10)
KAN	<i>rrs</i>	3 (4, 0)
	<i>eis promoter</i>	9 (10, 0)

First of all, two databases were consulted, *TBDreaMDB* (Sandgren *et al.* 2009) and *MUBII-TB-DB* (Flandrois *et al.* 2014). Lineage specific mutations and polymorphisms without sound phenotypic data supporting their association with resistance were discarded. In addition, recent literature was consulted to extract new DR mutations from review papers (Nebenzahl-Guimaraes *et al.* 2014; Lorenzo & Mousa 2011; Maruri *et al.* 2012; Georghiou *et al.* 2012), papers on TB DR tests (Liu *et al.* 2013; Moure *et al.* 2013; Shi *et al.* 2013; Wang *et al.* 2013; Zimenkov *et al.* 2013; Sekiguchi *et al.* 2007; Engström *et al.* 2012; Helb *et al.* 2010; Jin *et al.* 2012; Ajbani *et al.* 2012), papers on gene mechanisms of DR (Slayden & Barry 2000; Jagielski & Grzeszczuk 2013; Ando *et al.* 2010; Safi *et al.* 2013; Morlock & Metchock 2003; DeBarber *et al.* 2000; Brossier *et al.* 2011; Tan *et al.* 2013) and other recent studies (H. Zhang *et al.* 2013; Booniam *et al.* 2010; Jnawali *et al.* 2013; Lin *et al.* 2013). As presented in Table 4.1, the library comprised 1276 polymorphisms at 946 nucleotide positions from 25 loci, 6 promoters and 19 coding regions, involved in resistance to 11 drugs. In addition to examining individual drugs the cumulative loci for MDR and XDR-TB was considered. Circos software was used to construct circular genomic region variation maps (Krzyszowski *et al.* 2009). The R software package was used for statistical analysis.

4.2.2 Sequence data and drug susceptibility testing

The precision of the curated library for predicting resistance was assessed through analysis of new and published sequence data so that *in silico* inferred resistance phenotypes could be compared to phenotypes derived from conventional culture-based susceptibility studies. Only sample collections with raw sequencing data (minimum read length 50bp) and drug susceptibility data from recognized testing

protocols (Stop TB Partnership 2014) were considered. A total of 792 isolates from six geographically distinct data sets were used (WGS data set 3) as described in Section 1.7. Isolates were filtered as described in Section 2.2.1. Of the 792 isolates 365 (46%) were phenotypically resistant to at least one drug, 262 (33%) were MDR-TB, 54 (6.8%) XDR-TB and 426 (54%) were susceptible to the drugs tested. Two RMP mono resistant samples were reported and 99.2% of RMP resistant samples were MDR-TB. *In silico* genotyping using *SpolPred* (Section 3.2.2) revealed all major modern MTBC lineages were represented, including Lineage 1 (EAI family, n=68, 8.6 %), Lineage 2 (Beijing, n=182, 23 %), Lineage 3 (CAS, n=86, 10.9 %) and Lineage 4 (456 isolates, 57.5 %, of which 35 X, 97 T, 298 LAM, 4 S, 18 H, 4 other). Where conventional susceptibility data was not available samples were excluded from analysis for that drug. Sensitivity, specificity, accuracy, 95% confidence intervals (*CI*) for the statistical performance of each test (i.e. DR mutation list) were estimated using the following formula (Altman 1990):

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

where *TP* represents the number of true positives (phenotypic resistant isolates harbouring DR mutations), *FN* false negatives (phenotypic resistant isolates lacking DR mutations), *TN* true negatives (phenotypic susceptible isolates lacking DR mutations) and *FP* false positives (phenotypic susceptible isolates harbouring DR mutations). The

corresponding 95% confidence intervals for sensitivity and specificity were estimated using:

$$CI = p \pm z \sqrt{\frac{p(1-p)}{n}}$$

where p is either sensitivity or specificity, z is equal to 1.96 and n is the denominator in either the sensitivity ($n = TP + FN$) or specificity ($n = TN + FP$) formulas.

To compare the performance of a pair of genotypic tests (i.e. DR mutation lists), the difference between the underlying discordant results (i.e. number of samples being positive for test 1 and negative for test 2 (b) versus negative for test 1 and positive for test 2 (c)) was calculated. The statistical significance of this difference was determined using the common proportion difference test for binomial variables based on the following p-value:

$$p_{value} = 2 \left(1 - pnorm \left(\left| \frac{(b - c)}{\sqrt{(b + c)}} \right| \right) \right)$$


where $pnorm$ is the cumulative probability function of the standard normal distribution.

4.2.3 Rapid mutation detection and the *TB Profiler* Online tool

To rapidly characterise mutations from WGS files (*FASTQ* format), raw sequences were mapped to a modified version of the H37Rv reference genome using the *Snap* algorithm (Zaharia *et al.* 2011), and SNPs and indels of high quality called using *SAMtools/BCFtools* (Q30, 1 error per 1000bp) as previously described in Section 2.2.1.

(A)

TB Profiler

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE 

This tool processes raw sequence data to infer strain type and identify known drug resistance markers.

This tool is for **Research Use Only**. It has not been approved, cleared, or licensed by any regulatory authority. By submitting sequence data the user acknowledges no intended medical purpose or objective such as clinical diagnosis, patient management, or human clinical trials.

Results

The results for all jobs are available [here](#).

Submit

Please select one (single end) or two (paired end) gzipped FASTQ files to upload and process, each file must be under 1GB in size. If you choose to add a name for this analysis then do it carefully as it will be made public.

Public Name (optional):	<input type="text"/>
Gzipped FASTQ file:	<input type="button" value="Browse..."/> No file selected.
Second FASTQ (optional):	<input type="button" value="Browse..."/> No file selected.
<input type="button" value="Submit"/>	

The processing queue has 0 jobs in it.

Example data

Sample	FASTQ Data from the EBI	Profile
Malawi/Mixed/MDR	ERR176616	Profile
Malawi/Lineage 1	ERR190365	Profile
Malawi/Lineage 4-Style/Pan-Susceptible	ERR212132	Profile
China/Lineage 2-Beijing/XDR	SRR671726	Profile
Tibet/Lineage 4-T/XDR	SRR671740	Profile


Further Information

This tool, with application to a large, published dataset, is described in detail in the journal article:
 From genome to bedside: closing the gap with ultra-rapid analysis for tuberculosis drug resistance.
 F. Coll, R. McNerney, M.D. Preston, T.G. Clark, et al.
 <Submitted>

Please cite us if you use our tool.
 Processing time is under 10 minutes per sample plus queuing time; for example 2:30 minutes for a 500Mb file.

(B)

TB Profiler

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE 

This tool processes raw sequence data to infer strain type and identify known drug resistance markers.

This tool is for **Research Use Only**. Data and information provided through use of this tool are not intended for medical purpose or objective and should not be used for clinical diagnosis, patient management, or human clinical trials.

Back to [results](#) page.

Name: Malawi/Mixed/MDR
 Sample: ERR176616

Drug ¹	Resistance	Supporting Mutations
Isoniazid	R	katG (S315T)
Rifampicin	R	rpoB (S450L)
Ethambutol	R	embA (C-16T), embB (E378A), embC (T270I)
Pyrazinamide	R	pncA (Q141P)
Streptomycin	R	rs (-)
Ethionamide		
Fluoroquinolones		
Amikacin		
Capreomycin		
Kanamycin		
Multi drug resistance	R	
Extremely drug resistance		

Lineage ²	Name	Main Spoligotype	RDS
lineage1	Indo-Oceanic	EAI	RD239
lineage1.1	Indo-Oceanic	EAIQ,EAH4,EAIS,EAIG	RD239
lineage1.1.3	Indo-Oceanic	EAI6	RD239
lineage4.3	Euro-American (LAM)	mainly-LAM	None
lineage4.3.4	Euro-American (LAM)	LAM	RD174
lineage4.3.4.2	Euro-American (LAM)	LAM1;LAM4;LAM11	RD174
lineage4.3.4.2.1	Euro-American (LAM)	LAM11	RD174

Figure 4.1 The TB profiler tool

(A) Screenshot of TB profiler input page where FASTQ files and run Id are selected by the user (B) Screenshot of TB profiler output page with DR and lineage information (<http://tbdr.lshtm.ac.uk>).

The modified *Mtb* reference genome was created by concatenating the nucleotide sequence of all DR candidate loci (+/- 300 bp) plus positions harbouring strain-specific SNPs (+/- 700 bp). This resulted in a sequence of 187 kb, i.e. 4% of the original genome size. Subsequently, DR status and strain type are derived based on the presence of DR mutations in the curated list and strain-specific SNPs (Section 3.3.3) respectively. Other mutations (SNPs and small indels) in DR candidate genes not present in the curated list are also reported.

The online ***TB Profiler*** tool (<http://tbdr.lshtm.ac.uk/>) was developed in Perl/PHP. It inputs raw sequence data (FASTQ format), identifies DR and strain-specific mutations, and displays related outputs (see screenshots in Figure 4.1). A *perl* script was used to implement the *Snap* software and *SAMtools/BCFtools* based bioinformatic pipeline.

4.2.4 Comparison with existing tools

To examine the potential analytical advantage of WGS over current molecular technology for detecting DR, comparison was made with three commercial tests: (i) the Xpert MTB/RIF (Cepheid Inc, USA) which targets the *rpoB* gene for RMP^R; (ii) the LiPA MTBDRplus for MDR-TB (Hain Lifescience, Germany) which targets *rpoB*, *katG* and *inhA* for resistance to RMP and INH and (iii) the LiPA MTBDRsl (Hain Lifescience, Germany) which targets *gyrA*, *rrs* and *embB* for resistance to the FLQ, AMI and EMB respectively. Using the polymorphisms exploited within these assays (Helb *et al.* 2010; Jin *et al.* 2012; Ajbani *et al.* 2012), *in silico* versions were developed, and their performance was compared to the curated mutation library. In particular, *in silico* analysis of the six data sets was performed and analytical sensitivities and specificities

of the inferred resistance relative to the reported phenotype were compared (Figure 4.4 and Supplementary Figure 10). The *cumulative* effect on sensitivity and specificity of DR mutations were calculated for all drugs, MDR and XDR (Supplementary Figure 9).

4.3 RESULTS

4.3.1 Validation of mutation library

The mutation library was validated using new and publically available sequence and phenotypic data. *In silico* inferred resistance from WGS data was compared to the reported resistance phenotype from conventional culture-based testing. Results are summarised in Table 4.2.

Sensitivity and specificity of the whole genome analysis varied across drugs and with the geographic origin of the sample collections. For the drugs that contribute to MDR-TB correlation of mutation analysis with the reported phenotype was high.

Mutations predictive of resistance were found in 96.2% and 92.8% of samples resistant to RMP and INH, respectively. Of the 22 INH resistant samples predicted as susceptible, 14 were from China, 7 of those had mutations in known candidate loci (*katG* and *ahpC* promoter), which would explain resistance but were not previously reported (Table 4.3).

Table 4.2 Accuracy of whole genome drug resistance analysis compared to reported resistance phenotype

Drug (tested)	# R (%)	Sen (95%CI)	Spe (95%CI)	China Sen/Spe	Pakistan Sen/Spe	Malawi Sen/Spe	Portugal Sen/Spe	Russia Sen/Spe	Canada Sen/Spe
INH (693)	305 (44)	92.8 (89.9-95.7)	100 (100-100)	88/100	100/100	92.6/100	94.6/100	100/100	-/100
RMP (694)	264 (38)	96.2 (93.9-98.5)	98.1 (96.8-99.4)	95.7/97.7	97.3/100	100/98.2	96.9/100	90.9/90	-/100
EMB (484)	150 (31)	88.7 (83.6-93.8)	74.6 (69.9-79.3)	83.6/71.3	100/42.7	100/80	85.7/68.1	100/80	-/100
STR (487)	225 (46.2)	87.1 (82.7-91.5)	87.1 (86-93.4)	86.8/91	95.8/44.4	61.5/95.6	86.8/81.5	100/100	-/100
PZA (307)	110 (35.8)	70.9 (62.4-79.4)	93.9 (90.6-97.2)	NT	51.3/-	66.7/94.8	80.6/100	100/60	-/100
ETH (334)	155 (46.4)	73.6 (66.7-80.5)	93.3 (89.6-97)	38.9/97.3	66.7/90.3	NT	84.9/84.6	NT	NT
MOX (42)	10 (23.8)	60 (29.6-90.4)	68.7 (52.6-84.8)	NT	NT	NT	83.3/56.2	25/100	NT
OFX (313)	117 (37.4)	85.5 (79.1-91.9)	94.4 (91.2-97.6)	77.8/95.1	-/100	NT	92.1/93.2	NT	NT
AMK (193)	76 (39.4)	82.9 (74.4-91.4)	98.3 (96-100)	NT	86.5/100	NT	79.5/98.2	NT	NT
CAP (358)	89 (24.9)	60.7 (50.6-70.8)	90.7 (87.2-94.2)	50.0/97.0	85.7/21.7	NT	57.7/98.0	100/91.7	NT
KAN (316)	118 (37.3)	87.3 (81.3-93.3)	93.4 (89.9-96.9)	71.4/97.0	83.8/-	NT	98/88.7	80/33.3	NT
MDR (693)	262 (37.8)	91.2 (87.8-94.6)	98.4 (97.2-99.6)	86.3/100	97.3/100	100/98.2	95.8/100	90.9/90	-/100
XDR (601)	54 (9)	75.9 (64.5-87.3)	98.4 (97.3-99.5)	60.9/99.1	-/100	-/100	96.3/88.9	25/100	-/100

Abbreviations: NT, Not Tested; Sen, Sensitivity; Spe, Specificity

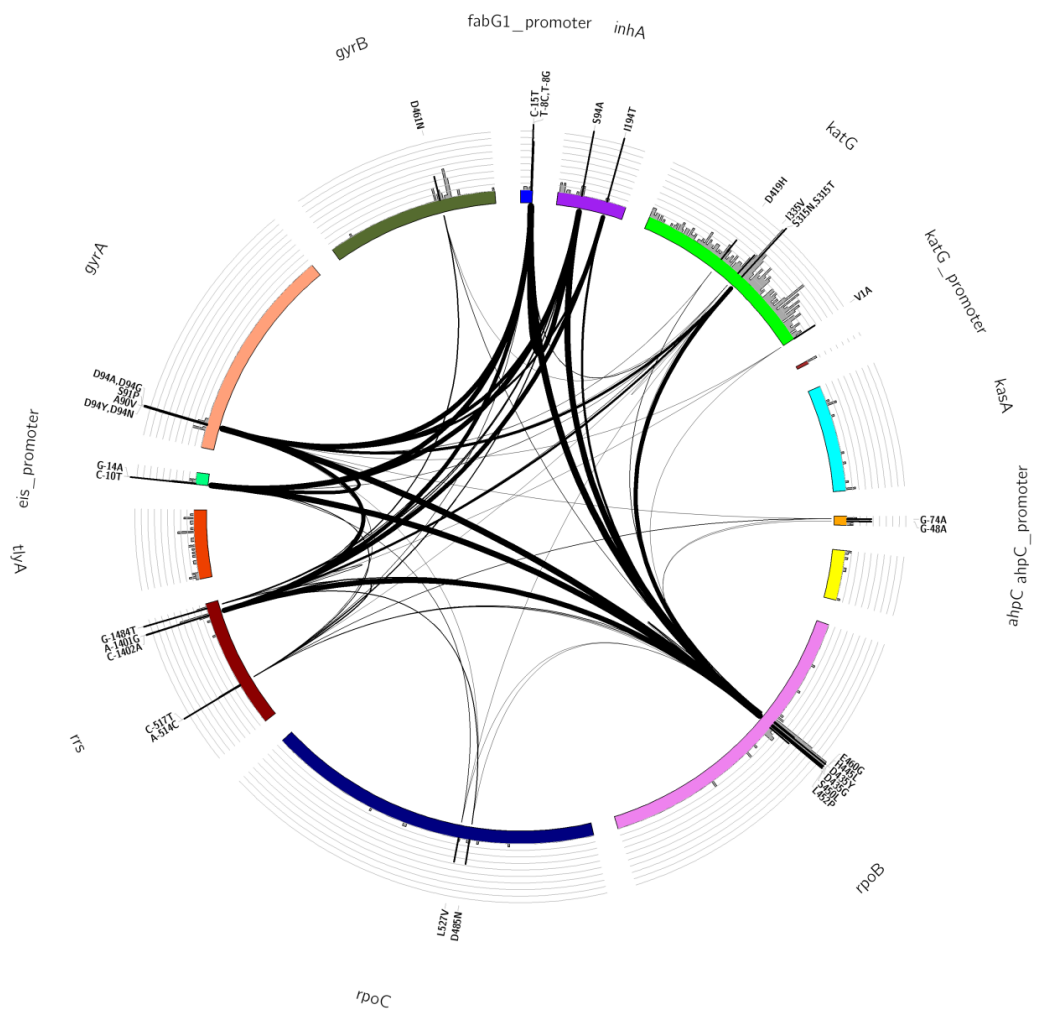


Figure 4.3 Mutations associated with XDR-TB found in phenotypically XDR strains

See footnote in Figure 4.2 for a description of this plot.

Correlation was lower for other first line drugs. For PZA, 32 of 110 samples with a resistant phenotype harboured no mutations in known DR genes, including 18 of 37 samples from Karachi. However, specificity for this drug was high (93%, 95% CI 90.6-97.2). Specificity for ETH was similarly high but accuracy was poor for EMB where 85 of 334 susceptible stains were found to harbour mutations included in the curated library of DR polymorphisms.

Among the AMI drugs, accuracy was higher for AMK (sensitivity/specificity 82.9/98.3%) and KAN (87.3/93.4%) than for CAP, where 35 of 89 resistant samples were not detected by the *in silico* genome analysis. Testing for FLQ resistance was less commonly reported and data for OFX was restricted to two studies (China and Portugal) with a total of 313 samples tested.

Ten OFX-susceptible samples were found to harbour mutations associated with FLQ resistance (94.4% specificity), and mutations were not identified in 17 resistant samples (85.5% sensitivity). Of 42 samples tested for susceptibility to MOX, 10 were reported as phenotypically resistant, of which 6 were recognized by the *in silico* mutation analysis (60% sensitivity). Figure 4.3 summarises the mutations and multiple loci associated with resistance to INH, RMP, FLQ and AMI and shows those found in phenotypically determined XDR cases. Supplementary Figure 8 illustrates the loci involved in resistance to each of the 11 drugs, the position of DR mutations in the curated list within these loci, and DR mutations observed in phenotypically resistance cases. The cumulative effect of these DR mutations on sensitivity and specificity for each drug is shown in Supplementary Figure 9.

4.3.2 Comparison with commercial tests

Having assessed the diagnostic potential of the mutation library comparisons were made with polymorphisms used in commercial molecular tests for DR. Results are summarised in Figure 4.4. When screening for resistance to RMP there was no significant difference between the performance of curated library and mutations employed by the Xpert MTB/RIF and the LiPA MTBDRplus. However, 31 samples had

mutations predictive of resistance to INH not included in the line probe assay, resulting in a 10% drop in sensitivity for MTBDRplus.

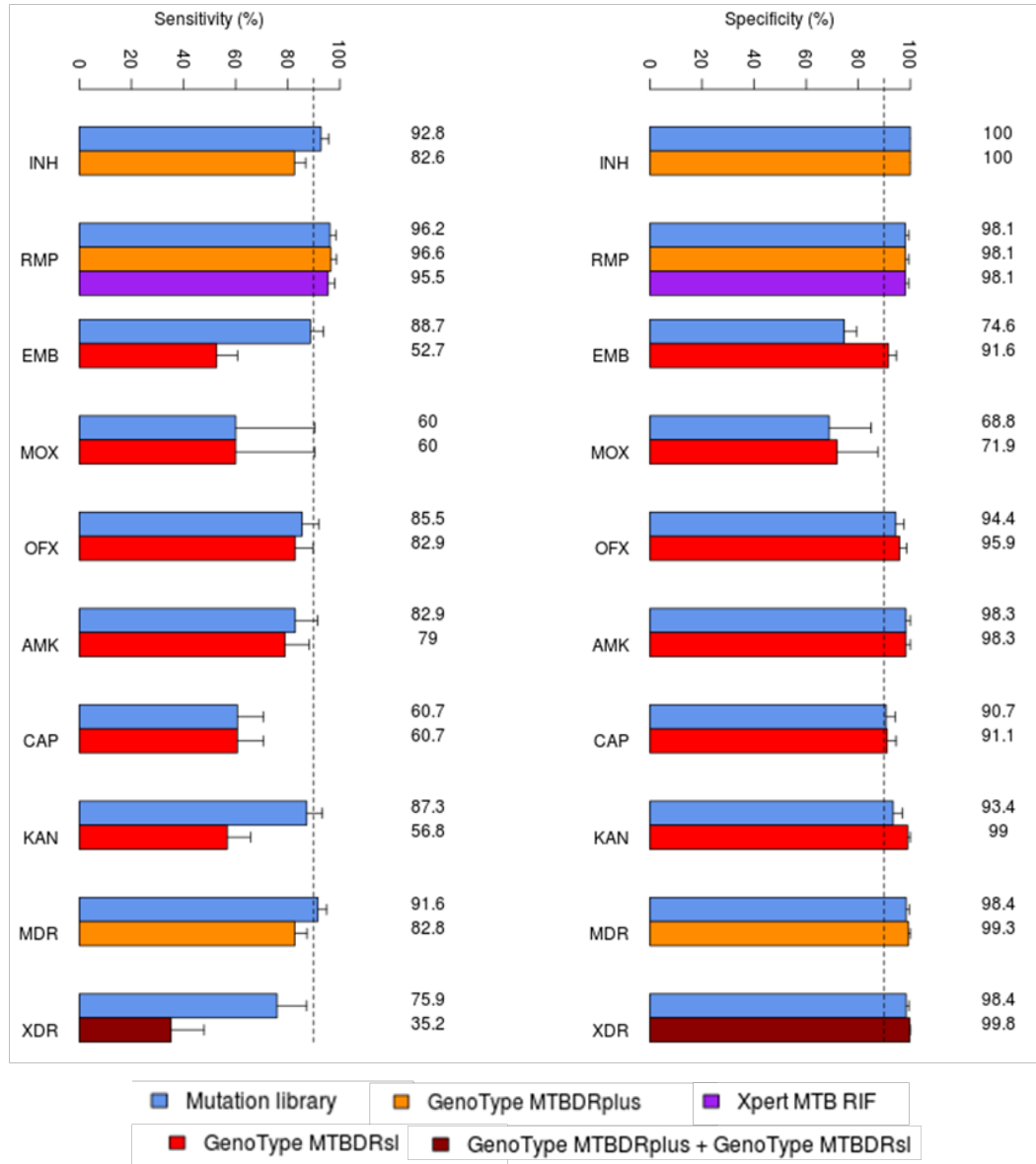


Figure 4.4 Inferred analytical accuracies of the DR curated mutation library and three commercial molecular tests for resistance

In silico analysis of published sequence data using mutation libraries derived from MTBDRplus, MTBDRsl, Xpert (Cepheid Inc, USA) MTB/RIF (Hain Life Sciences, Germany) and the curated whole genome library. For each library, *in silico* inferred resistance phenotypes were compared to reported phenotypes obtained from conventional DST. Sensitivity and specificity percentages are accompanied with 95% confidence intervals.

The mutations concerned were mainly in the *katG* gene: S315N (n=9), S315G (n=1), D419H (n=1), L378P (n=1), V1A (n=1), Y155C (n=3), W191R (n=5 and always with C-15T *inhA* promoter), N138D (n=1, with T-8A *inhA* promoter) and T380I (n = 1, with C-15T *inhA* promoter). There were also 6 samples with *ahpC* promoter mutations and 2 samples with *inhA* mutations (S94A and I194T). No resistance mutations were observed in INH susceptible strains (i.e. 100% specificity).

The curated library offered enhanced accuracy over the line probe mutations when screening for MDR-TB (95.8 vs. 93.1 %; $p < 0.00042$). Detection of resistance to EMB, OFX, AMK and KAN was also enhanced by the whole genome analysis. A slight reduction in specificity was observed for five drugs: EMB (91.6 vs. 74.6%, $p < 1.54e-08$), MOX (71.9 vs. 68.8%, $p < 0.32$), OFX (95.9 vs. 94.4%, $p < 0.083$), CAP (91.1 vs. 90.7%, $p < 0.32$), KAN (99.0 vs. 93.4%, $p < 0.00091$). Less susceptibility data was available for the second line drugs. For each of the FLQ and AMI the sensitivity of the curated DR library was equal to, or greater than for the mutations employed in the LiPA MTBDRsl: MOX (60% in both cases), OFX (85.4 vs. 82.9%, $p < 0.083$), AMK (82.9 vs. 78.9%, $p < 0.083$), CAP (60.7% in both cases) and KAN (87.3 vs. 56.8%, $p < 1.97e-09$). Overall when detecting XDR-TB resistant cases the curated mutation library offered enhanced accuracy over the line probe mutations (96.3 vs 93.7%; $p < 0.0047$) (Figure 4.4).

4.3.3 Comparison with other drug resistance databases

The diagnostic accuracy of other DR mutation databases, namely *TBDreaMDB* (Sandgren *et al.* 2009) and *MUBII-TB-DB* (Flandrois *et al.* 2014), was estimated and compared to that obtained by the herein presented library. Accuracy for the detection of RMP resistance did not differ significantly among the three databases (Figure 4.5).

INH specificity was considerably lower (23.4%) when using *TBDreaMDB* markers, mainly due to the presence of lineage-specific mutations, which have been historically and mistakenly regarded as resistance associated mutations (e.g. R463L *katG* for non-lineage 4 strains).

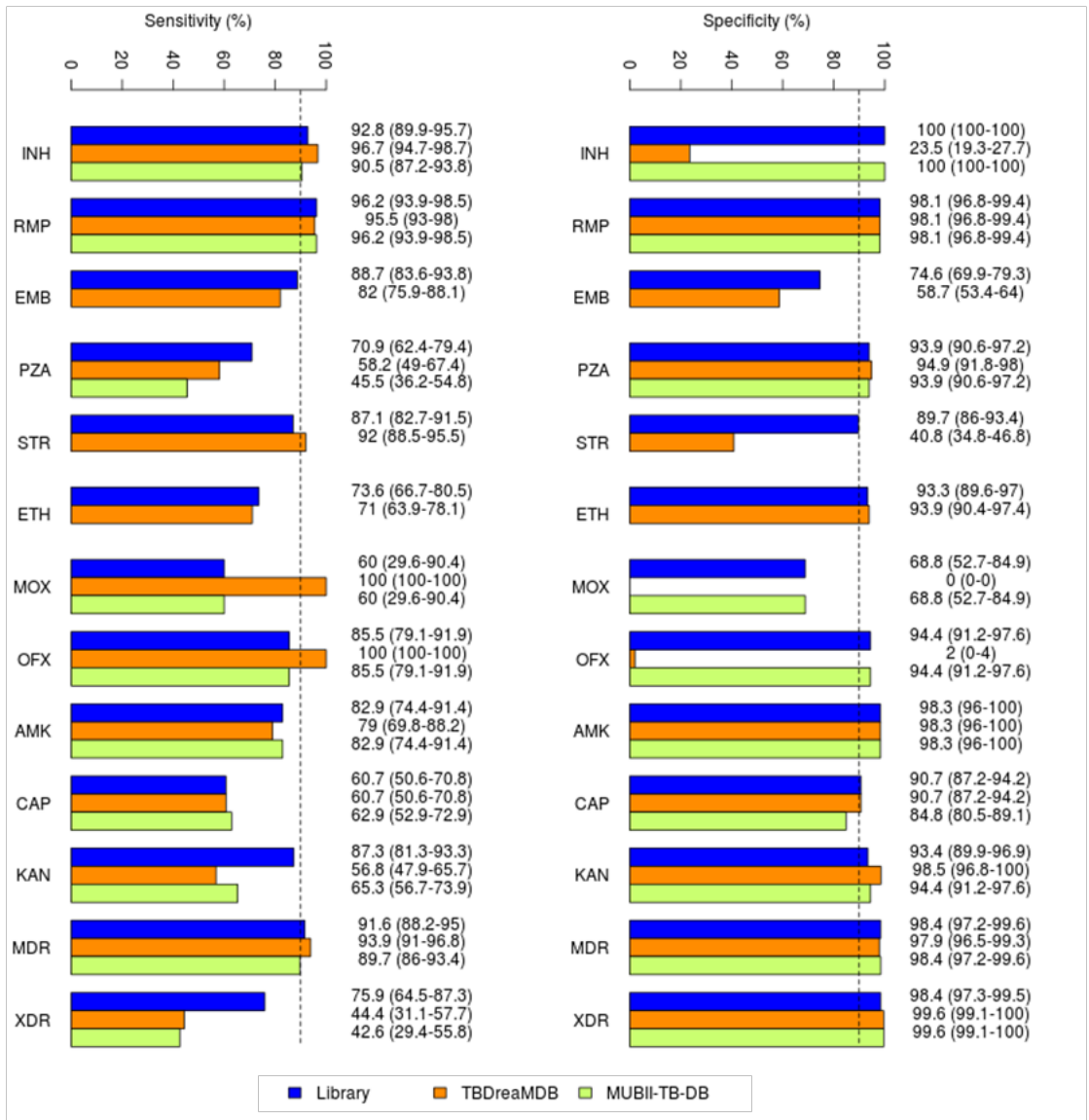


Figure 4.5 Diagnostic performance of the curated library versus alternative drug resistance mutation databases

In silico analysis of published sequence data using mutation libraries derived from *TBDreaMDB*, *MUBII-TB-DB* and the curated library. For each library, *in silico* inferred resistance phenotypes were compared to reported phenotypes obtained from conventional DST. Sensitivity and specificity percentages are accompanied by 95% confidence intervals (See Methods).

Slightly more INH resistant cases could be correctly assigned (283/305, 92.8% sensitivity) than *MUBII-TB-DB* (276/305, 90.5% sensitivity) without compromising specificity (100% in both cases), and consequently more MDR cases too, 240/262 (91.6%) and 235/262 (89.7%) respectively. The diagnostic accuracy for other first-line drugs (PZA and STR), not included in current molecular diagnostic tests, also improved (Figure 4.5).

Of 110 PZA resistant samples, 78 (70.9% sensitivity) were detected compared to 50 (45.5%) and 64 (58.2%) identified by *MUBII-TB-DB* and *TBDreaMDB* respectively. While specificity was relatively high for both ETH (93.3%) and PZA (93.9%), not all resistant cases were found to harbour mutations in known genes (51 and 25 respectively). Resistant clinical isolates lacking mutations in candidate genes have been extensively reported for both ETH (Brossier *et al.* 2011) and PZA (Stoffels *et al.* 2012a). STR specificity by *TBDreaMDB* was significantly low (40.8%) due to the presence of lineage-specific SNPs in the *gid* gene (e.g. E92D for Modern Beijing and L16R for LAM), which were found in both STR susceptible and resistant cases and therefore unlikely to be involved in STR resistance.

Both *MUBII-TB-DB* and the curated library produced similar predictive performance for most of second line drugs (OFX, MOX, AMK and CAP). *TBDreaMDB* reports several mutations (E21Q, G668D and S95T in *gyrA*) that do not correlate with FLQ resistance and were responsible for the very low specificity for FLQ resistance (Maruri *et al.* 2012). By including mutations at the *eis* promoter, 103 of 118 KAN resistant samples were identified (87.3% sensitivity) compared to 67 (56.8%) and 77 (65.2%) predicted by *TBDreaMDB* and *MUBII-TB-DB* respectively. Overall, the presented database

outperformed both *TBDreaMDB* (75.9% vs. 44.4%, $p < 0.00021$) and *MUBII-TB-DB* (75.9% and vs. 42.6% $p < 2.21 \times 10^{-5}$) in the detection of XDR cases.

4.3.4 Online tool for predicting drug resistance and lineage information from sequenced isolates

Having established a curated list of mutations for eleven anti-TB drugs, a web-based tool to rapidly identify a DST and strain-type profile was developed. In addition, to identifying known mutations, this research tool can identify novel mutations in the candidate regions, thereby leading to further functional characterisation and update to the curated list.

Table 4.3 Potentially novel DR mutations identified by *TB profiler*

Drug	DR Candidate Gene	Mutations*	Number of samples (Population)	Increased sensitivity
INH	<i>katG</i>	G/299/S, P/232/S, F/408/L, D/142/G, G/120/S, CTCGGGT/2155245/C, D/189/A, D/419/G	7 (China)	2.3%
	<i>ahpC</i> promoter	C/2726136/T		
RMP	<i>rpoB</i>	S/450/Stop, CAGCCAGCTG/761087/C	2 (Karachi and Portugal)	0.7%
EMB	<i>embA</i>	C/4243225/T, G/4243190/C, C/4243218/CTACCATCGAG	7 (6 from China, 1 from Portugal)	4.7%
	<i>embB</i>	G/554/D, G/200/S, A/679/T, Y/319/D, S/538/P, S/412/P, N/399/T		
PZA	<i>pncA</i>	V/130/M, G/2289011/GT, I/133/S, G/2288786/GGCCAAGCCAT (n=2), G/2289011/GT	7 (6 from Karachi, 1 from Portugal)	6.4%
	<i>rpsA</i>	Q/410/R (n=2)		
STR	-	-	-	-
ETH	<i>fabG1</i> promoter	T/1673432/G, T/1673432/A	9 samples (4 from China, 5 from Portugal)	5.7%
	<i>inhA</i>	I/95/L, I/194/T		
	<i>ethA</i>	CT/4326393/C, GT/4327132/G, A/4326800/AGC, C/403/R (n=2), Y/143/Stop, P/51/S, P/149/S		

*Mutations not present in susceptible cases, which are not strain-specific SNPs or synonymous SNPs and therefore more likely to be conferring DR. SNPs in coding regions are annotated with the reference amino acid, codon number and alternative amino acid. SNPs in non-coding regions and indels are annotated using the reference nucleotide allele, chromosome coordinate and alternative allele as extracted from the VCFs.

This approach called *TB profiler* aligns raw sequencing data to a small version of the *Mtb* reference genome, consisting of 4% of the original chromosome size and containing only DR candidate loci and regions harbouring strain-specific SNPs. *TB profiler* processed *FASTQ* files at a rate of 80,000 sequence reads per second. Application to the 792 samples led to the identification of 36 novel mutations (24 nsSNPs, 9 indels and 5 intergenic SNPs) in phenotypically resistant strains and absent in susceptible ones, which could be potentially driving resistance (Table 4.3). The online tool is available from <http://tbdr.lshtm.ac.uk/> (Figure 4.1).

4.4 DISCUSSION

The global emergence and amplification of resistance to anti-TB drugs has created a need for improved detection tools. Conventional DST, which in this work is assumed to be the reference standard, requires several weeks to complete and variation in both protocols and MIC standardization can lead to inconsistent results. Therefore, point-of-care diagnostic tests for rapid detection of all available anti-TB drugs are urgently needed to guide treatment options for patients with MDR, XDR-TB and post XDR (TDR-TB) disease. The potential of WGS to provide such a solution was assessed. A library of approximately 1,300 mutations to predict resistance to eleven drugs was assembled. This library has been incorporated into a rapid online tool to perform the analysis and provide a DST and strain-type profiles. The presented library is the most complete and updated database of its kind and gathers the state of the art on our understanding of the genetic basis of DR in TB.

In situ analysis of sequence data was undertaken to validate the library. Sensitivity was highest for RMP. The mode of action of this drug (Section 4.1.2) involves only one gene

and is fully understood compared other TB drugs that have more complex modes of action. Unsurprisingly the sensitivity was lower for drugs like PZA and ETH, for which the understanding of the genetic basis of resistance is currently incomplete (Brossier *et al.* 2011; Stoffels *et al.* 2012a). Further work is needed to determine additional polymorphisms predictive of resistance to these drugs. As illustrated in Figure 4.2 and Figure 4.3 a large number of loci are associated with MDR and XDR.

A limiting factor for this study is the reliability of culture-based susceptibility testing methods and the lack of a gold standard with which to compare new tests. Previous studies on discrepancies between mutation and culture derived phenotypes suggest that molecular assessment may eventually become the gold standard for some drugs (Rigouts *et al.* 2013; Van Deun *et al.* 2013).

In addition to assessing additional numbers of drugs, the sensitivity of the whole genome mutation library was equal to, or greater than the mutations used in the commercial LiPAs for all drugs examined, demonstrating the intrinsic advantage of a whole genome approach over current LiPA tests, which include a limited number of DR mutations. It was also found more accurate than previously reported databases, due to increased numbers of polymorphisms strongly predictive of resistance and the absence of mutations with weak or no supporting data, mainly strain-specific mutations. The enhanced sensitivity was greatest for INH, KAN and for MDR and XDR-TB. Specificity for RMP, INH, AMK, MDR and XDR-TB exceeded 98%. Results for EMB were less promising as although a sensitivity of 88.5% was achieved, the specificity of 71.3% is inadequate. These results are in line with different levels of EMB resistance being acquired through mutations in multiple loci, some of which are currently unknown

(Safi *et al.* 2013). Although the current state of knowledge does not allow EMB resistance to be predicted at high precision, EMB resistant mutation in the curated list can be helpful at pointing to strains with higher predisposition to develop such resistance. It should be noted that high positive predictive value is crucial for DR tests where the consequence of a false positive may be unnecessary isolation in specialist containment facilities and prolonged treatment with drugs of increased toxicity.

The accuracy of the mutation analysis was observed to vary by geographic region (Supplementary Figure 10). The isolates used in this dataset (WGS data set 3) were not necessarily representative of the local population and geographic disparities in the frequency of DR mutations may reflect the clonal nature of TB transmission, which is entirely human-to-human. It has been suggested that emergence of resistance in *Mtb* is associated with bacterial lineage and difference in the prevalence of polymorphisms could relate to variance in *Mtb* lineage across regions but such conclusions cannot be drawn from the present study as more appropriate sampling strategies are required.

Accuracy values also differed among drugs belonging to the same group. MOX resistance was predicted using the same markers as OFX (FLQ markers) yielding overall worse specificity (68.7 vs. 94.4%). There is considerable cross-resistance between FLQ but MIC values can vary among members of this group. MOX generally presents lower levels of resistance than OFX for the same mutations (Maruri *et al.* 2012), which would explain its lower specificity. Strains having the same FLQ resistance-conferring mutations are more likely to be regarded as sensitive (false positives) for MOX. Sensitivity for MOX also needs to be improved (60%). The low number of MOX-resistant cases (n=10) may be biasing the calculated sensitivity. Similarly, CAP

specificity was lower than AMK (90.7 vs. 98.3%) and KAN (90.7% vs. 93.4%). 24 of the 269 CAP-susceptible strains had the A1401G *rrs* SNP, a frequently reported AMI-resistance conferring mutation, also found in 50/89 CAP-resistant samples. The finding of AMI-resistant mutations in CAP-susceptible samples (i.e. low specificity) has already been observed and explained in terms of a high MIC cut-off recommended by the WHO for CAP (Rodwell *et al.* 2013).

Not all drugs used in the treatment of TB were included in this study. Drugs were omitted either because insufficient susceptibility data was available (e.g. CIP and rifabutin (RFB)), or because the mechanism of action remains elusive and SNPs to predict resistance have yet to be identified (e.g. cycloserine (CYS) and para-aminosalicylic acid (PAS)).

Recent work has shown that pyrosequencing has sufficient sensitivity to test DR susceptibility in *Mtb* clinical specimens, thereby significantly shortening the turnaround time for obtaining molecular DST results to hours (Lin *et al.* 2013). These assays focus on detecting only the most prevalent mutations within short sequences (<50 bases) in limited numbers of genes. In contrast, WGS technologies allows interrogation over all genes with sequencing of much longer segments, making it possible to identify mutations spread across the locus as is the case in *pncA* gene. Low frequency mutations, which may be predominant in certain geographical areas and complex mechanisms of resistance involving multiple loci (e.g. EMB) are also accessible using WGS. Potentially new DR-conferring mutations in candidate loci can also be investigated as demonstrated in Section 4.3.4. Furthermore, newly identified DR loci and mutations can easily be incorporated in the mutation library. Such flexibility will

allow new drugs to be included, for example bedaquiline, which has recently been found to have cross resistance with clofazimine through the mutations in the transcriptional regulator *Rv0687* (Hartkoorn *et al.* 2014).

Rapid WGS may not only replace current methods for identifying and typing MTBC, but also has also the potential to detect DR TB (Sharon J. Peacock 2013). Phenotypic susceptibility testing cannot still be replaced for all antibiotics since the genetic mechanisms of resistance are not fully characterised. Future work is needed to identify new loci closely associated with resistance to drugs like PZA, STR, ETH, EMB, MOX and CAP, and recently licensed anti-TB drugs. Nevertheless, WGS can be used to rapidly identify resistance when mutations proven to confer resistance are detected and provide valuable clinical information to guide treatment.



Registry

T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

COVER SHEET FOR EACH 'RESEARCH PAPER' INCLUDED IN A RESEARCH THESIS

Please be aware that one cover sheet must be completed for each 'Research Paper' included in a thesis.

1. For a 'research paper' already published

- 1.1. Where was the work published?
- 1.2. When was the work published?
 - 1.2.1. If the work was published prior to registration for your research degree, give a brief rationale for its inclusion
.....
.....
.....
- 1.3. Was the work subject to academic peer review?
- 1.4. Have you retained the copyright for the work? **Yes / No**
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 copyright holder (publisher or other author) to include work

2. For a 'research paper' prepared for publication but not yet published

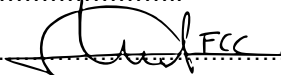
- 2.1. Where is the work intended to be published? Nature genetics
- 2.2. Please list the paper's authors in the intended authorship order
See next page
- 2.3. Stage of publication – Not yet submitted / Submitted / ~~Undergoing revision from peer reviewers' comments~~ / ~~In press~~


3. 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)

See next page

NAME IN FULL (Block Capitals) FRANCESC COLL I CEREZO

STUDENT ID NO: 323873

CANDIDATE'S SIGNATURE  Date 10/12/2014

SUPERVISOR/SENIOR AUTHOR'S SIGNATURE (3 above) 

RESEARCH PAPER 5

PAPER DETAILS: Francesc Coll, Grant A. Hill-Cawthorne, Kim Mallard, Rumina Hasan, Zahra Hasan, Nerges Mistry, Rob Warren, Keertan Dheda, Patricia Sheen, David Moore, Jaime Robledo, Maxine Caws, Stefan Pantaiotov, Richard Anthony, Saad Alghamdi, Joao Perdigao, Miguel Viveiros, Isabel Portugal, Andy Ramsey, Bouke de Jong, Leen Rigouts, Theolis Bessa, Tomoshige Matsumoto, Anabela Miranda, Noram Mocillo, Christophe Sola, Ruth McNerney, Arnab Pain, Taane G Clark. A whole genome association approach reveals insights into global Mycobacterium tuberculosis drug resistance.

AUTHORS CONTRIBUTION:

FC conducted all analyses under the guidance of TGC. GA-CH, KM, and AP coordinated the sequencing. RH, ZH, NM, RW, KD, PS, DM, JR, MC, SP, RA, SA, JP, MV, IP, AR, BdJ, LR, TB, TM, AM, NM and CS contributed DNA samples and meta data, including strain-typing and drug susceptibility testing data. RM, AP, and TGC are joint PIs on the project.

Chapter 5

Identification of novel drug
resistance associated loci
using genome-wide
association approaches

5 IDENTIFICATION OF NOVEL DRUG RESISTANCE ASSOCIATED LOCI USING GENOME-WIDE ASSOCIATION ANALYSIS

A complete characterisation of the genetic determinants of DR is a prerequisite for accurate genetic-based DSTs. As well as being a potential DR detection tool as shown in the previous chapter, WGS can be used to characterise new genes involved in DR. In this regard, this chapter explores the use of WGS to capture the natural genetic variants of drug resistant and susceptible clinical strains and assess the association of these with phenotypic DR.

5.1 INTRODUCTION

The emergence and spread of antimicrobial DR is an enormous public health concern. Dissecting the genetic determinants of antibiotic resistance has important implications in both diagnosis of resistance and development of effective alternative therapies. The mutations found to be associated with DR can be incorporated into genotypic drug susceptibility assays and facilitate tailored treatment (Lacoma *et al.* 2008). On the other hand, the genes harbouring these mutations can provide insights into the bacterial mechanisms that underlie DR and assist in the rational design of novel antimicrobial agents (Lee *et al.* 2014).

Early attempts aimed at identifying DR mechanisms in *Mtb* consisted in studying intrinsic drug susceptibility in other mycobacterial species (Mdluli *et al.* 1998; Danilchanka *et al.* 2008). High-throughput mutagenesis approaches have been used to discover mutated loci responsible for changes in drug tolerance, followed by complementation with intact versions of these loci that restored the initial DR

phenotype (Danilchanka *et al.* 2008; Ren & Liu 2006). *In vitro* drug exposure and subsequent isolation of surviving resistant mutants, a process known as directed evolution, has also been used to identify the genetic changes responsible for DR acquisition (Hartkoorn *et al.* 2014; Safi *et al.* 2013). These studies have characterised multiple DR acquisition pathways, including the role of cell-wall permeability and the involvement of efflux pumps. Once DR genes are identified, allelic exchange experiments have been extremely useful at establishing causality and elucidating the mutation's contribution to the DR phenotype (Nebenzahl-Guimaraes *et al.* 2014). Although extremely successful, *in vitro* studies do not necessarily capture the genetic causes of DR observed in clinical specimens. WGS enables the full characterisation of naturally occurring genetic variation in clinical isolates which, coupled with an accurate drug phenotype characterisation, can be a powerful approach to dissect the genetic determinants of DR and other clinically important phenotypes.

Recent studies have applied genome-wide approaches to systematically search for genes closely related to DR in *Mtb* (H. Zhang *et al.* 2013; S. Zhang *et al.* 2013; Farhat *et al.* 2013; Casali *et al.* 2014). These studies employed WGS to genotype a panel of MTBC clinical isolates with different resistance profiles. Subsequently, loci associated with DR can be identified by genome sequence comparisons between susceptible and resistant isolates. This approach allowed the identification of a new possible mechanism of PZA resistance (S. Zhang *et al.* 2013). Genome-wide association analysis (GWAS) is a powerful approach established in human disease (Anon 2007), with established statistical techniques, which aims to measure the statistical significance of phenotype-genotype associations. Since DR is generally a binary categorical variable (coded as

resistant or susceptible), GWAS make use of a logistic regression framework, which can include adjustment for confounders such as pre-existing DR or strain-type (population structure). A recent study in 161 *Mtb* clinical samples (H. Zhang *et al.* 2013) across 8 drugs demonstrated the value of the GWAS approach. Whilst identifying established DR loci, the multi-genic nature of some associations highlighted that the genetic basis of DR may be more complex than previously anticipated. Despite the identification of a set of new loci, the specific role of these genes in DR was not discussed. GWAS have also been applied to dissect the genetic causes of DR (Alam *et al.* 2014) and virulence phenotypes in other bacterial pathogens (Laabei *et al.* 2014).

In parallel, a novel method to identify genetic markers of resistance was recently developed (Farhat *et al.* 2013) and applied to MTBC clinical strains with different antibiograms. This method consists of sequencing the genomes of related strains with different resistance phenotypes followed by a phylogenetic-based genome-wide scan for positive selection. This approach (*phyC*) assumes that genetic variants associated with DR are under convergent positive selection and therefore originate *de novo* across independent lineages (i.e. branches of the phylogenetic tree). This test has shown to be valuable as it identified well known DR markers in MTBC. The involvement of drug efflux pumps and DNA repair genes in DR acquisition was also highlighted.

In addition to association studies, WGS can also shed light on the molecular mechanisms underlying the transmissibility and persistence of DR strains in a population. It is well known that DR acquired mutations may infer a fitness cost in the absence of antibiotic pressure (de Vos *et al.* 2013). Compensatory mutations may then arise and restore the fitness of resistant bacteria. This insight has important

epidemiological consequences as MTBC DR clinical strains harbouring fitness-compensatory mutations have been associated with higher transmissibility and persistence in the population (de Vos *et al.* 2013; Casali *et al.* 2014).

A collection of 2765 MTBC isolates from 18 global populations and different antibiograms (27.7% MDR, 17.1% XDR) has been used to identify novel mechanisms of DR. Two complementary approaches were applied to this dataset: tree-based convergent evolution (Farhat *et al.* 2013) and GWAS. Previously known resistance loci were identified by the *phyC* analysis as well as PE/PPE genes and other potential targets of convergent positive selection. The GWAS analysis identified all known drug-targets and drug-converting enzymes, in addition to transporters, loci involved in synthesis and regulation of cell wall components and genes of unknown function.

5.2 METHODS

5.2.1 Dataset, raw sequence alignment and SNP calling

A global dataset consisting of 2,902 MTBC clinical samples was compiled across multiple populations from different geographical areas and with representation of all main four lineages (1 to 4) (Table 5.1). Some of these datasets were downloaded from the public domain when both WGS and phenotypic drug susceptibility data were made publicly available. For the remaining ones, both the phenotypic and WGS data were provided by collaborators (see WGS data set 3 description in Section 1.7).

The percentage of DR varied across populations and was particularly low in the Vancouver and Karonga populations, which were included with the aim of having a

good representation of pan-susceptible samples with diverse genetic backgrounds (i.e. different lineages) in the final dataset.

Table 5.1 Summary of the global drug resistance dataset

Population	N	N (Post QC)	Mean DOC	# SNPs	Proportion of lineages	% of any R	% of MDR	% of XDR
Brazil	108	106	110	7606	4 (100)	100	100	10.38
Bulgaria**	17	17	155	6329	2 (5.88), 3 (5.88), 4 (82.35), <i>M. bovis</i> (5.88)	ND	ND	ND
China (H. Zhang <i>et al.</i> 2013)*	161	130	106	16537	2 (70.77), 3 (1.54), 4 (26.92)	73.08	73.08	4.62
Colombia	15	15	111	2764	4 (100)	100	100	0
Vancouver (Gardy <i>et al.</i> 2011)*	36	33	55	1035	4 (100)	0	0	0
India	17	15	59	5878	1 (26.67), 2 (6.67), 3 (40), 4 (26.67)	13.33	13.33	0
Japan**	4	4	283	1603	4 (100)	ND	ND	ND
Karachi	42	42	485	7681	1 (11.90), 3 (78.57), 4 (9.52)	88.1	88.1	0
Karonga (Malawi)	1662	1642	99	43064	1 (16.50), 2 (4.08), 3 (11.69), 4 (67.42), mixed (0.18)	7.31	0.61	0
Lisbon (Perdigão <i>et al.</i> 2013)	84	70	206	6518	2 (7.14), 3 (1.43), 4 (91.43)	72.86	54.29	1.43
Netherlands**	14	14	389	992	4 (100)	ND	ND	ND
Peru	104	99	290	8700	2 (6.06), 4 (93.94)	61.62	27.27	0
Porto	131	128	292	8152	2 (11.72), 4 (88.28)	84.38	39.84	6.25
Russia (Casali & Nikolayevskyy 2012)*	42	23	36	2524	2 (100)	65.22	52.17	8.7
South Africa	174	172	188	15478	1 (0.58), 2 (31.98), 3 (5.81), 4 (61.63)	98.26	48.84	1.16
Kampala, Uganda**	51	51	256	8019	1 (1.96), 2 (1.96), 3 (27.45), 4 (68.63)	ND	ND	ND
Vietnam	50	47	192	9256	1 (36.17), 2 (44.68), 4 (19.15)	53.19	40.43	0
WHO-TDR**	190	157	123	21689	1 (9.55), 2 (23.57), 3 (3.18), 4 (63.69)	ND	ND	ND
Overall	2902	2765	136	107462	1 (10.7), 2 (11.6), 3 (9.4), 4 (66.6), <i>M. bovis</i> (0.04) and mixed (1.7)	27.70	17.09	1.03

* Publicly available datasets ** Datasets with missing or unreliable phenotypic data

Overall 27.7% of the samples were resistant to at least one of the 11 drug tested (AMK, CAP, EMB, ETH, INH, KAN, MOX, OFX, PZA, RMP and STR), 17.1% of samples were MDR

(i.e. resistant to both INH and RMP), and 1% were XDR-TB (MDR in addition to resistance to any FLQ and AMI). It should be noted that each sample was tested for a different number of drugs, where susceptibility to first-line treatments usually led to no tests for second-line drugs.

The sequence data analysis procedures used are explained in Section 2.2.1. In brief, after mapping the raw sequence data to the H37Rv reference genome, SNPs were called and retained if present in unique regions of the genome resulting in 107,462 SNP sites. Samples were removed if they had more than 15% of SNP missing calls (n=2,765/2,902).

5.2.2 Phylogenetic reconstruction and population structure

The best-scoring maximum likelihood phylogenetic tree was computed using *RAxML* (Stamatakis *et al.* 2008) based on the 107,462 SNP sites spanning the whole genome and the resulting tree rooted on *M. canettii* as described before (Section 3.2.4). All isolates were *in silico* genotyped using *SpolPred* (Section 3.2.2). The proposed SNP typing system described in Section 3.3.3 was employed to accurately genotype all isolates at both lineage and sub-lineage levels. In addition, a principal components analysis (PCA) was conducted to capture the population structure, as well as adjust for it in the GWAS analysis.

5.2.3 Phylogenetic convergence test for selection

To identify SNPs enriched by convergent evolution, the *phyC* approach described in (Farhat *et al.* 2013) was employed using the available implementation in (Alam *et al.* 2014). Briefly, the constructed *RAxML* tree and the whole-genome SNP alignment were

used as input to perform ancestral sequence reconstruction using available functions in the R *phangorn* module (Schliep 2011). Mutations specifically enriched in branches leading to resistant leaf nodes of the tree (i.e. samples) compared to susceptible ones were identified and statistical significance measured using a Fisher's exact test (Alam *et al.* 2014).

5.2.4 Genome-wide association analysis

DR for each drug was a binary response variable (0 for susceptible, 1 for resistant, NA or missing for non-determined). Logistic regression was used to estimate the strength of association between each locus in the genome and resistance to each drug. SNP mutations were aggregated by coding region, RNA loci and intergenic regions resulting in a matrix of $n \times m$, where n is the number of loci and m the number of samples. For coding regions only non-synonymous mutation changes were aggregated. In addition, SNPs were grouped by operons, functional units containing clusters of genes under the control of the same promoter and generally involved in the same pathway or function. Operon annotation was extracted from TBDB (Reddy *et al.* 2009).

Since resistance to anti-TB drugs originates on top of pre-existing DR, as a result of current TB treatment regimens, the sequence of resistance accumulation was estimated based on associations among resistance phenotypes. Logistic regression was then performed adjusting for pre-existing DR as well as population structure, using the first five principal components. Specifically the models fitted were of the form:

$$\log \frac{P}{1-P} = \alpha + \beta_1 SNP + \beta_2 Previous + \beta_3 PC1 + \beta_4 PC2 + \beta_5 PC3 + \beta_6 PC4 + \beta_7 PC5,$$

where P is the probability of resistance, $\log(P/1-P)$ refers to the log odds, SNP refers to the number of mutations at the locus to be tested, $Previous$ refers to the results for other drugs tested, $PC1 - PC5$ are the principal components, and the alpha and betas refer to log odds of intercept and the corresponding variables respectively. For example, the model for estimating the association of resistance to PZA of *Rv2043c* (*pncA*) used the number of nsSNPs for each sample within *Rv2043c* (SNP), INH, RMP, STR, and EMB as binary variables indicating the results of these tests for samples ($Previous$). Statistical significance of each variable was established using a Wald test. Of most interest was the SNP effect, with its odds ratio and p-value. The standard errors for the log odds ratios were estimated from the model, and used to construct 95% confidence intervals for the odds ratios.

5.3 RESULTS

5.3.1 Population structure

Genome analysis performed on 2,765 MTBC clinical isolates from 18 independent populations revealed substantial genetic diversity, with 107,462 SNP sites in non-repetitive regions of the genome relative to the H37Rv reference strain. Figure 5.1 shows the phylogenetic tree constructed using all 107k SNPs. In addition, samples were classified using the 62 strain-specific SNP system (Section 3.3.3) and colour-coded accordingly on the genome-wide phylogeny, and demonstrated perfect clustering of lineage and sub-lineage groups as expected.

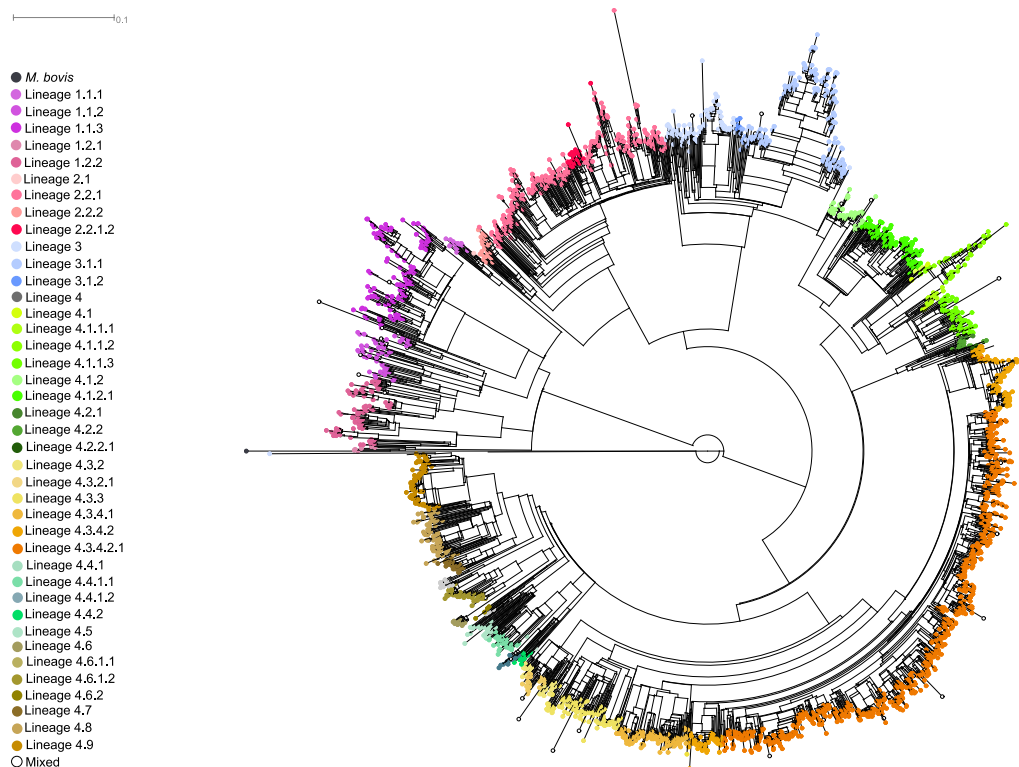


Figure 5.1 Phylogeny of the global drug resistance MTBC samples

Maximum likelihood phylogenetic tree constructed with 2,765 MTBC samples and rooted on *M. canettii*. The tips of the tree (i.e. samples) are colour-coded by lineage and sub-lineage based on the SNP typing system presented in Section 3.3.3.

A total of 2,699 out of 2,765 samples (97.6%) were unambiguously classified at both the lineage and sub-lineage levels, whereas a minority of 46 isolates (1.7%) presented markers from multiple groups. These samples belonged predominantly to the Karonga dataset (39/44) and harboured combinations of SNPs specific to different lineages (e.g. 1 and 4) or different sub-lineages of the same lineage (e.g. 4.3 and 4.6). The fact that Karonga district in Malawi is an area of high TB prevalence and that these SNP patterns are indeed combinations of markers from the most frequent circulating strain-types in this population (Guerra-Assunção *et al.* 2014) supports the hypothesis of mixed infection cases (Supplementary Table 9). A total of 20 samples from lineage4 (0.7%) did not harbour sub-lineage specific SNPs.

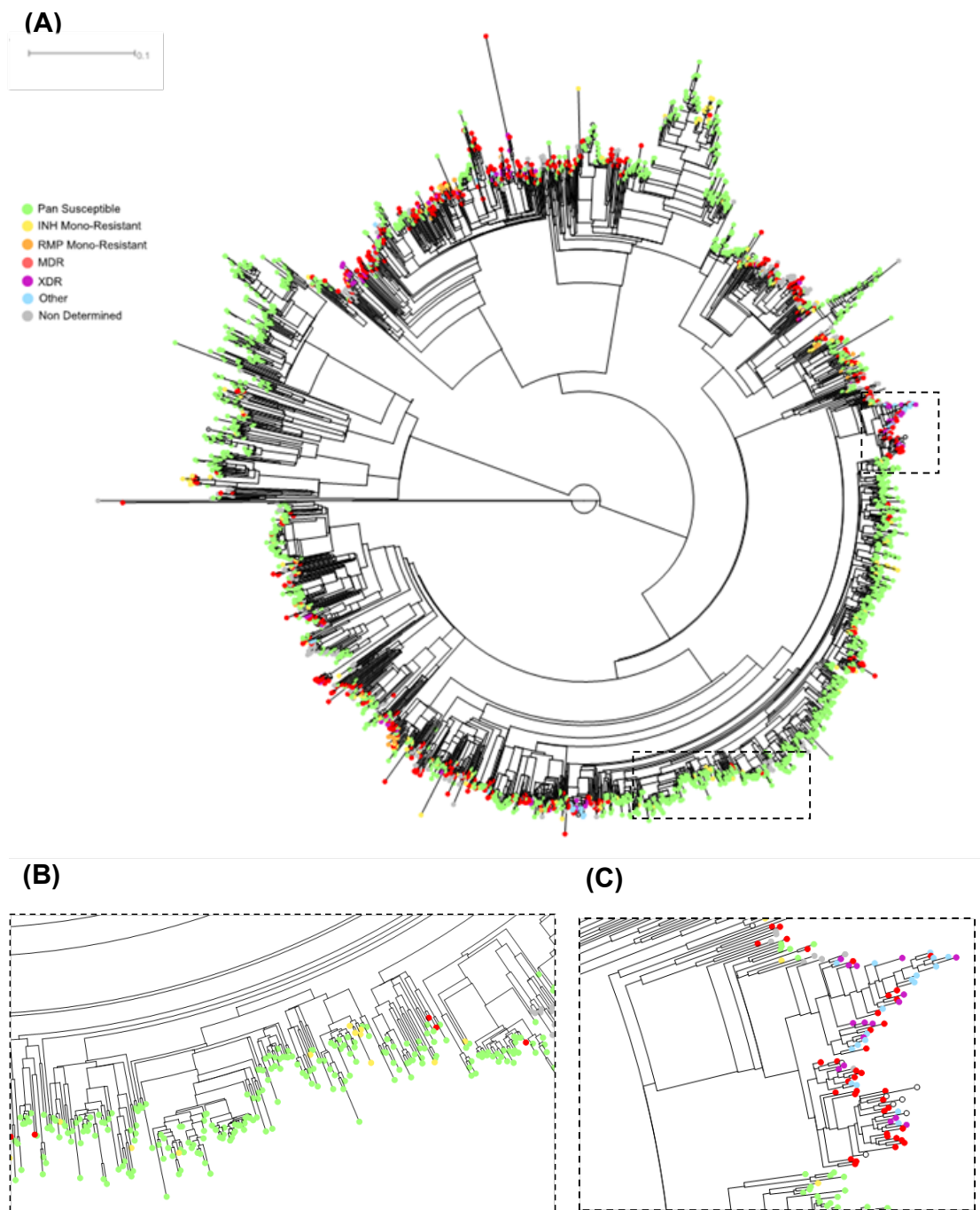


Figure 5.2 Phylogeny of the global drug resistance MTBC samples colour-coded by drug resistance status

(A) Maximum likelihood phylogenetic tree constructed with 2,765 MTBC samples and rooted on *M. canettii*. The tips of the tree (i.e. samples) are colour-coded by DR status. (B) Apparent absence of DR transmission in samples from Karonga, Malawi (B) Plausible scenario of DR transmission among samples from Lisbon and Porto populations.

Follow-up investigation of these samples revealed that they belong to four potentially novel sub-lineages consisting of 5, 5, 8 and 2 samples respectively. In total, and after excluding potential mixed samples (n=46, 1.7%), representatives of lineage1 (n=296, 10.7%) and all main three modern MTBC lineages (2, n=322, 11.6%; 3, n=259, 9.4%; 4, n=1842, 66.6%) were reported.

In order to highlight possible cases of DR transmission, samples in the phylogenetic tree were colour-coded by DR status (Figure 5.2A), namely as pan-susceptible (sample susceptible to all drugs for which it was tested), INH mono-resistance, RMP mono-resistance, MDR and XDR. Overall, 27.7% of samples were at least resistant to one anti-TB drug, 17.1% were MDR and 1% XDR. Figure 5.2C shows a plausible scenario of ongoing DR transmission among samples from Lisbon (Perdigão *et al.* 2013) and Porto populations, whereas Figure 5.2B illustrates the development of DR in multiple strains independently and an apparent absence of DR transmission in samples from the Karonga, Malawi, population.

5.3.2 Phenotypic drug resistance explained by known candidate genes

It was first determined whether phenotypic resistance could be explained by mutations in previously described DR-associated genes. The compiled library of putative DR mutations (Section 4.2.1) was used to infer an *in silico* DR phenotype from the genomic data. These “inferred” phenotypes were then compared to those from conventional culture-based DST, assuming the latter are the reference standard. Samples without available DST phenotypes were excluded from the analysis. Figure 5.3 shows that the percentage of explained resistance varied per drug, with resistance to first line drugs (INH, RMP and EMB) being genetically explained in a greater number of

samples, with the exception of PZA (60.3%) and STR (70.9%). Genetically determined resistance to second line drugs (ETH, OFX, AMK and CAP) was commonly lower, with the exception of OFX (83.3%) and KAN (82.7%). Having found that some drugs were not explained well by the mutation library, the focus shifted to drug-resistant samples lacking known DR-associated genetic markers, and the presence of unreported mutations in candidate genes (see Table 4.1 in Section 4.2.1 for a complete list of DR candidate genes).

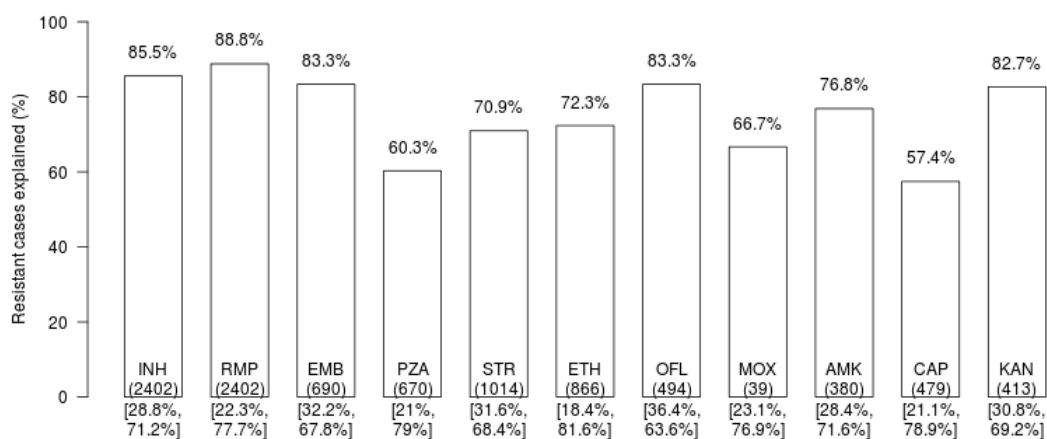


Figure 5.3 Percentage of resistance cases explained by known DR markers

The percentage of phenotypically resistant cases harbouring known DR mutations is indicated at the top of each bar. The number of samples tested is indicated in parenthesis at the bottom of each bar, as well as the proportion of resistant and susceptible cases among tested samples (enclosed in square brackets).

For each drug, Table 5.2 includes the number of DR samples lacking known DR mutations (column 2), and the subset of these (column 5) harbouring mutations in candidate genes that could potentially be the cause of resistance (column 3).

For example, of the 100 INH^R isolates without known resistance markers, 10 harboured mutations in *katG* (4 indels and 9 nsSNPs) that can potentially explain the resistant phenotype. Strain specific mutations, sSNPs and mutations present in susceptible

samples were not regarded as potentially DR-conferring mutations. Still, 46 potentially novel ones (35 nsSNPs + 11 indels) could be added to the library of DR mutations (Section 4.2.1) currently containing 1276 SNPs and indels. These mutations would explain an extra 1.5, 0.7, 4, 5.7, 4.4 and 1.1% of phenotypic resistance for INH, RMP, EMB, PZA, ETH and OFX respectively.

Table 5.2 Previously unreported mutations in candidate genes

Drug	N ^a (%)	Non-strain specific mutations ^b	Strain specific SNPs	N ^c
INH	100 (14.5)	F657L_katG, TC2156070T_katG, L159F_katG, G120S_katG, CTCGGGT2155245C_katG, D189A_katG, D419G_katG, GCGC2155747G_katG, P232S_katG, F408L_katG, T667P_katG, TCG2156104T_katG, D142G_katG	R463L_katG, G2726105A_ahpCpromoter	10
RMP	60 (11.2)	GACCAGA761109G_rpoB, S450*_rpoB, CAGCCAGCTG761087C_rpoB, V113I_rpoB	A1075A_rpoB, G876G_rpoB, A542A_rpoC, G594E_rpoC, R173R_rpoC, A172V_rpoC, P601L_rpoC	4
EMB	37 (16.7)	N399T_embB, G200S_embA, S538P_embB, L304L_embB, S412P_embB, T546I_embB, Q445R_embB, G4243190C_embApr promoter, C4243218CTACCATCGAG_embApr promoter, A679T_embB, G554D_embA	C76C_embA, V981L_embC, R927R_embC, C4243225T_embApr promoter, A1092A_embA	9
PZA	56 (39.7)	V130M_pncA, S164*_pncA, I133S_pncA, C2289056CT_pncA, D12G_pncA, C2289136CCAGGTAGTCGCTG_pncA, V260I_rpsA, Q410R_rpsA	R212R_rpsA	8
STR	96 (29.1)	-	-	0
ETH	44 (27.7)	T1673432G_fabG1promoter, CT4326393C_ethA, C403R_ethA, I95L_inhA, P51S_ethA, I194T_inhA, P149S_ethA, A4326800AGC_ethA	-	7
OFX	30	R448H_gyrA, R592S_gyrA	G668D_gyrA, E21Q_gyrA, S95T_gyrA	2
MOX	3 (33.3)	-	G668D_gyrA, E21Q_gyrA, S95T_gyrA	0
AMK	25 (23.2)	-	-	0
CAP	43 (42.6)	-	C1472337T_rrs, L11L_tlyA	0
KAN	22 (17.3)	-	-	0

^a Phenotypically resistant samples without known DR-associated mutations. ^b Potentially new DR-associated mutations, found in resistant strains and absent in susceptible ones. Strain-specific SNPs and sSNPs were discarded. ^c Number of phenotypically resistant samples harbouring new DR-associated mutations, i.e. potential new cases of genetically explained DR.

GWAS results

Some resistant strains lack mutations in known DR genes (encoding for protein targets of the drug or drug-metabolizing enzymes), highlighting that the genetic mechanisms of resistance are not fully characterised, or that some of the phenotypes may not be robust. In order to improve the understanding of the genetic basis of DR in *Mtb* a GWAS was used to identify novel loci closely associated with resistance.

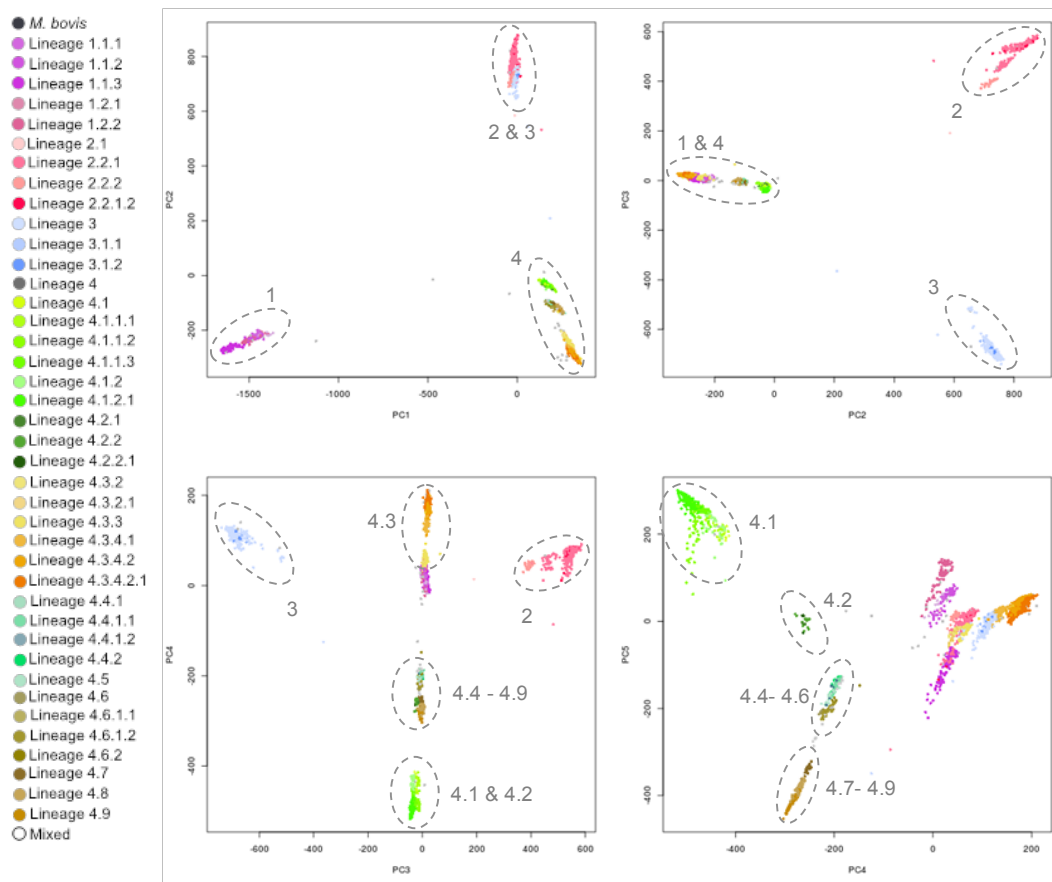


Figure 5.4 Principal Components Analysis for the global drug resistance data set

The 2,765 MTBC samples in the global drug resistance data set are plotted along PCs 1 to 5. Samples are colour-coded by lineage and sub-lineage based on the SNP typing system presented in Section 3.3.3. The first three components separate samples by main lineages (1 to 4) while PCs 4 and 5 provide sub-lineage separation.

As the population structure may confound the outcome of a GWAS, a PCA was conducted. The first three principal components (PCs) captured 73.2 % of the variance and as expected they distinguished the four main lineages (1 to 4) unambiguously. PCs 4 and 5 provided further sub-division at the sub-lineage level, particularly within lineage 4. These five PCs captured 82.1 % of the variance all together and were used directly as covariates in the logistic regression models (Figure 5.4).

Another expected confounder is the background DR, namely the existence of resistance to other drugs other than the one being tested for association. Figure 5.5A and Figure 5.5C show the GWAS results for INH before and after adjusting for overlapping resistance respectively. Figure 5.5B shows the number of INH^R samples being also resistant to other drugs. Well-known DR loci (*katG*, *inhA*, *fabG1* promoter, *rpoB*, *embB*, *pncA* and *rpsL*) were found to be strongly associated with INH^R (Figure 5.5A). The strength of association of genes not involved in INH^R dropped significantly after adjusting for background resistance (Figure 5.5C), while *katG* and *fabG1* promoter, known to be associated with INH^R, remained high. These results exemplify the confounding effects of background resistance and highlight the need of adjusting for it.

In addition to considering coding and intergenic regions independently, the GWAS was additionally performed at the operon level. As genes in the same operon are co-transcribed and generally participate in the same function, certain DR acquisition pathways may involve mutations in multiple genes (including the promoter) of the same operon. By aggregating SNPs by these functional units (e.g. genes or operons), the association signal could be boosted.

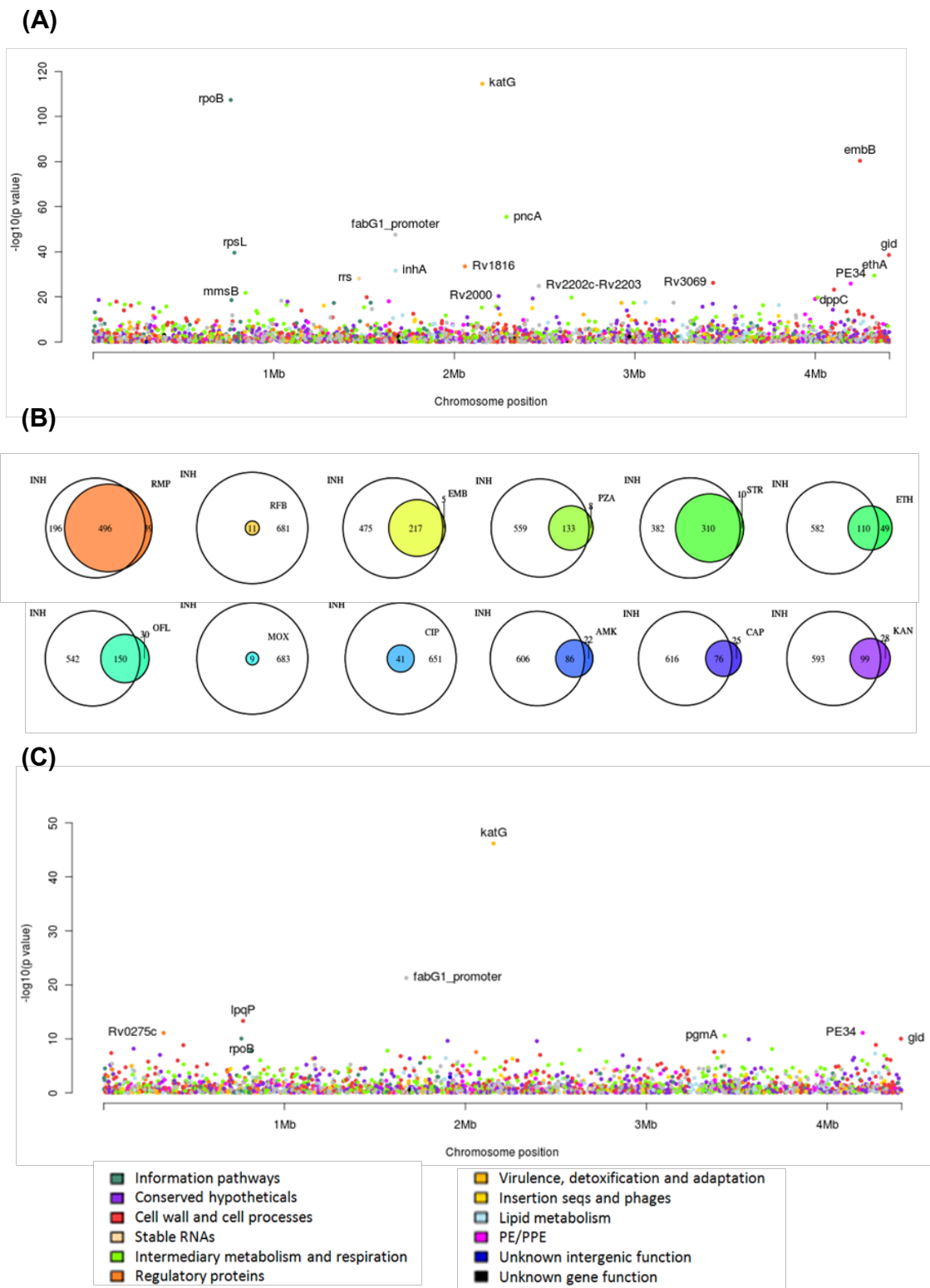


Figure 5.5 Locus GWAS results for isoniazid

(A) INH GWAS performed for both coding and intergenic regions. The y-axis is the log₁₀ p-value. For example, a value of 2 refers to a p-value of 0.01. Loci are colour-coded by functional category and plotted along the chromosome (x-axis). (B) Venn diagrams represent the number of samples being resistant to each pair of drugs. (C) INH GWAS for both coding and intergenic regions after adjusting for overlapping DR.

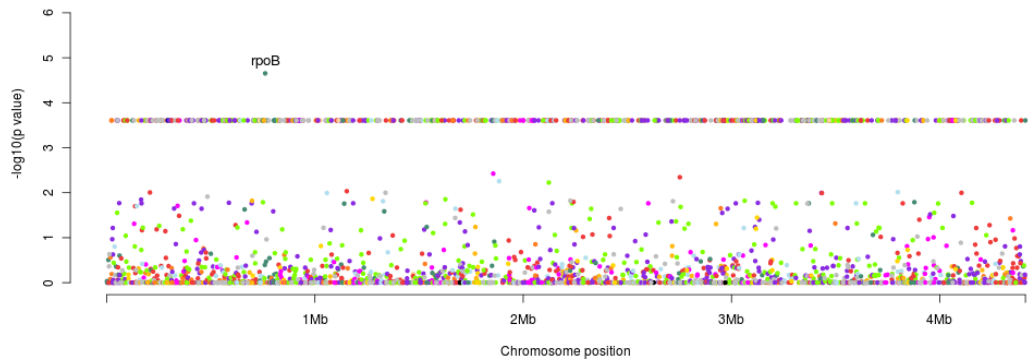


Figure 5.6 Locus GWAS results for rifampicin
See footnote in Figure 5.5 for a description of this plot

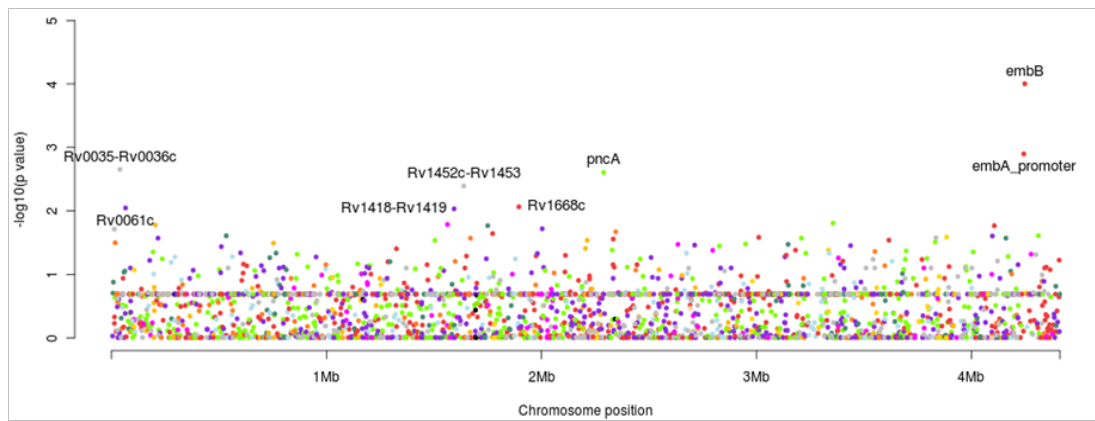


Figure 5.7 Locus GWAS results for ethambutol
See footnote in Figure 5.5 for a description of this plot

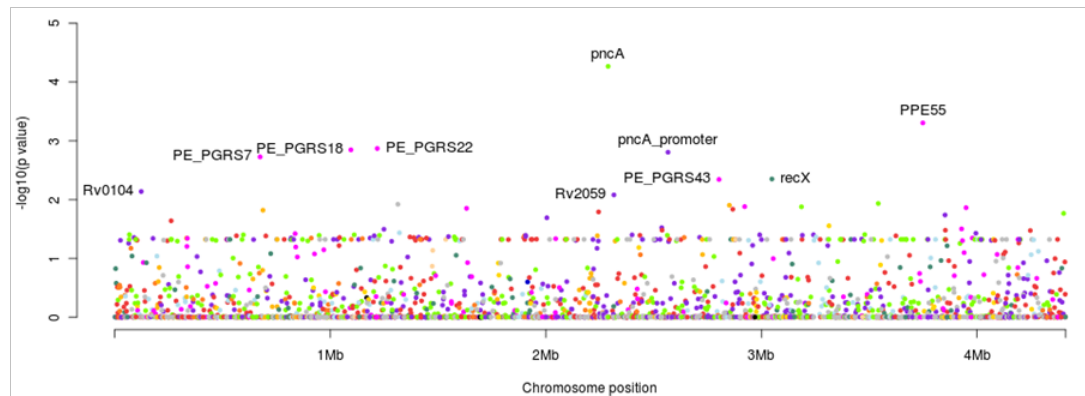


Figure 5.8 Locus GWAS results for pyrazinamide
See footnote in Figure 5.5 for a description of this plot

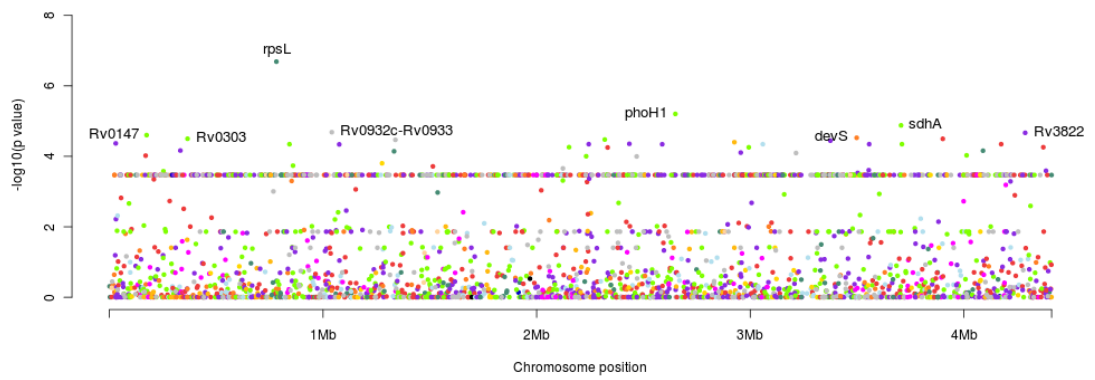


Figure 5.9 Locus GWAS results for streptomycin
See footnote in Figure 5.5 for a description of this plot

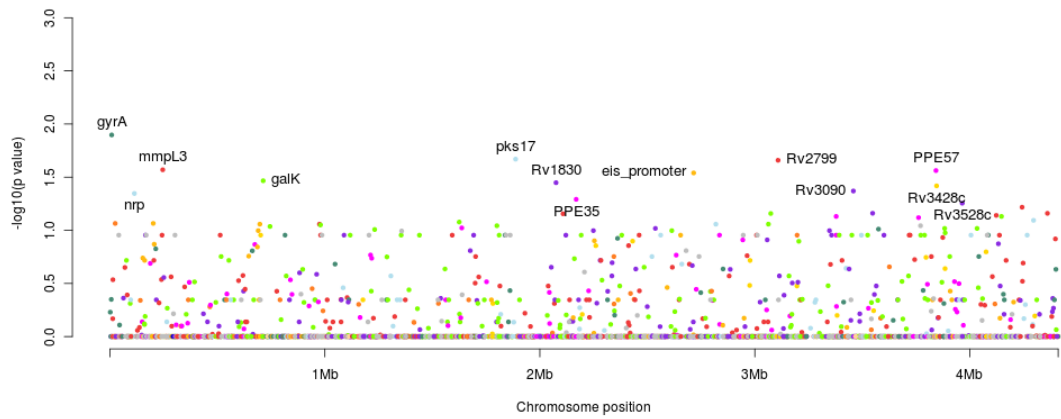


Figure 5.10 Locus GWAS results for ofloxacin
See footnote in Figure 5.5 for a description of this plot

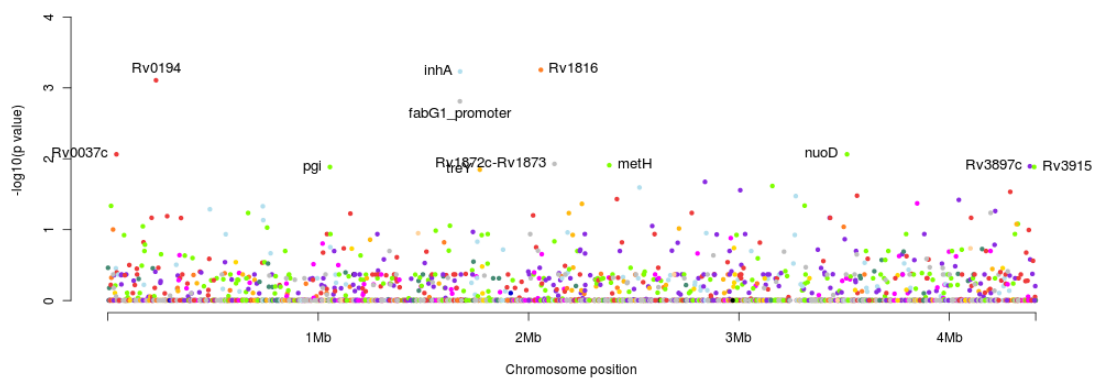


Figure 5.11 Locus GWAS results for ethionamide
See footnote in Figure 5.5 for a description of this plot

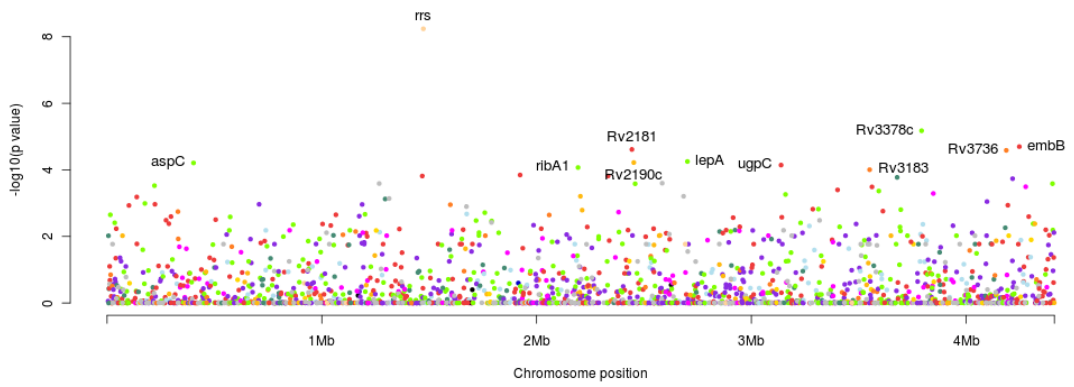


Figure 5.12 Locus GWAS results for amikacin

See footnote in Figure 5.5 for a description of this plot

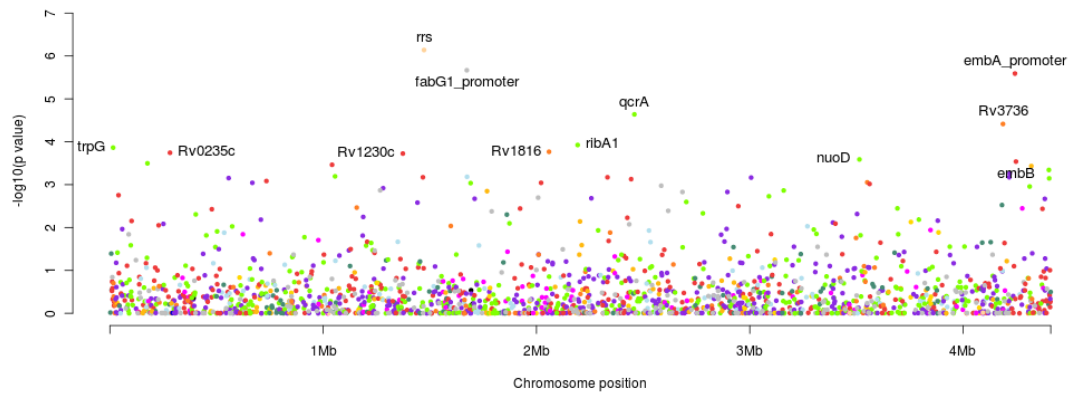


Figure 5.13 Locus GWAS results for capreomycin

See footnote in Figure 5.5 for a description of this plot

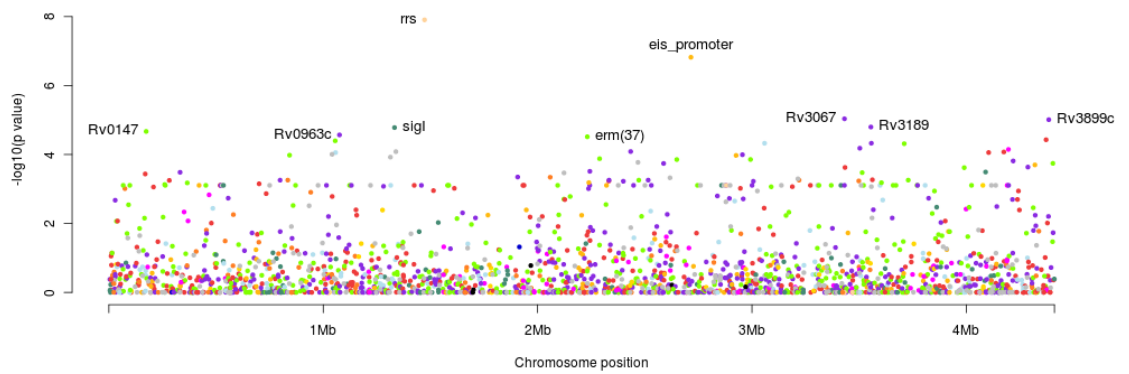


Figure 5.14 Locus GWAS results for kanamycin

See footnote in Figure 5.5 for a description of this plot

The majority of well-known DR loci were identified (Figure 5.5-5.14) which demonstrates the robustness of the GWAS approach and its implementation in this study. Loci (i.e. coding and intergenic regions) are colour-coded by functional category (see legend in Figure 5.5) in the Manhattan plots. There were a few notable absences though, including *rpsA* (Shi *et al.* 2011) and *panD* (S. Zhang *et al.* 2013), two recently proposed PZA^R-associated genes, as well as *rrs* (absent in the STR GWAS), *ethA* (ETH^R-associated gene), *rpoC* (RMP^R-associated) and *gyrB* (FLQ^R-associated). In addition to the classical DR genes encoding for drug-targets and drug-converting enzymes, a group of highly and intermediately DR-associated loci were also identified. Characterised transporters (e.g. *dppC*, *Rv1668c*, *mmpL3*, *ugcC* and *Rv0194*), probable transmembrane proteins (e.g. *Rv3061*, *Rv3069*, *Rv0143c*, *Rv0037c*, *Rv2799* and *Rv0235c*), PE/PPE genes (e.g. *PE34*, *PPE55*, *PPE57*, *PE_PGRS28*, *PE_PGRS7*, *PE_PGRS43* and *PE_PGRS18*) and other cell wall-related loci were particularly overrepresented within this group (Supplementary Table 10). The role of *Rv0194* in ETH^R deserves greater attention given its strong association with ETH^R and its documented function as multidrug efflux pump in *Mtb* (Danilchanka *et al.* 2008). The unexpected detection of putative transcriptional regulatory proteins (*Rv1816*, *Rv3736*, *Rv0275c* and *Rv2736c*) is an intriguing finding as they may underlie more complex mechanisms of DR. Indeed, these transcriptional regulators were found to be intermediately associated with resistance to multiple drugs (Supplementary Table 10) with the exception *Rv1816*, being strongly associated with ETH^R (Figure 5.11).

The operon-based GWAS (Supplementary Table 11) yielded results comparable with the locus-based ones, as operons containing known DR genes showed the highest level

of association. For instance, the *fabG1-hemZ* operon (which includes *fabG1* promoter and *inhA*) was closely related to INH and ETH resistance as expected, *Rv1907c-furA* (containing *katG* and *katG* promoter) to INH, *embA-embB* to EMB, *Rv2037c-pncA* to PZA and *rpsL-rpsG* to STR. In addition to these expected hits, the operon GWAS revealed potentially new players involved in DR including the *ugpC-ugpA* operon, which encodes for an ABC transporter; *Rv1635c*, possibly involved in the biosynthesis of lipoarabinomannan; *pstB-pstS1*, an ABC-type lipoprotein transporter; and other functionally uncharacterised operons (*Rv2295*, *Rv2203*, *Rv3528c*, *ccrB-Rv3071*, *Rv2313c-Rv2315*, *Rv3067*). These results demonstrate the usefulness of an operon-based GWAS approach to discover functionally-related regions associated with DR.

***PhyC* results**

If hits established using association methods are also under selective “convergent evolution” pressure, then this provides strong evidence of their role in resistance. The phylogenetic converge test identified the vast majority of known resistance determinants (Figure 5.15, in red; Supplementary Table 10, grey background rows); recently described DR-associated genes: *folC* (Zhao *et al.* 2014) and *ubiA* (Safi *et al.* 2013); and other targets of independent mutation (TIM) in both coding and intergenic regions. A sizeable number of TIMs mapped to the large family of PE/PPE genes (Figure 5.15, in blue). A few cases of genes involved in cell wall biosynthesis or remodelling were found (*lppB*, *lprP*, *mmpL12*, *pks7*, *pks12* and *pks15*). Despite the proven usefulness of the *phyC* test for uncovering genetic determinants of DR, there are several inherent limitations to this approach (Farhat *et al.* 2013). First, the detection power drops significantly for drugs with high proportion of missing (i.e. non-

determined) cases. This may result in statistically non-significant results for second-line drugs. Secondly, the background resistance confounding effects cannot be adjusted, meaning that drug-specific mutations are challenging to dissect in the presence of close associations among resistance phenotypes. The strength of the *PhyC* approach compared to GWAS, is the detection of other potential targets of convergent evolution due to selective pressures other than antibiotic exposure.

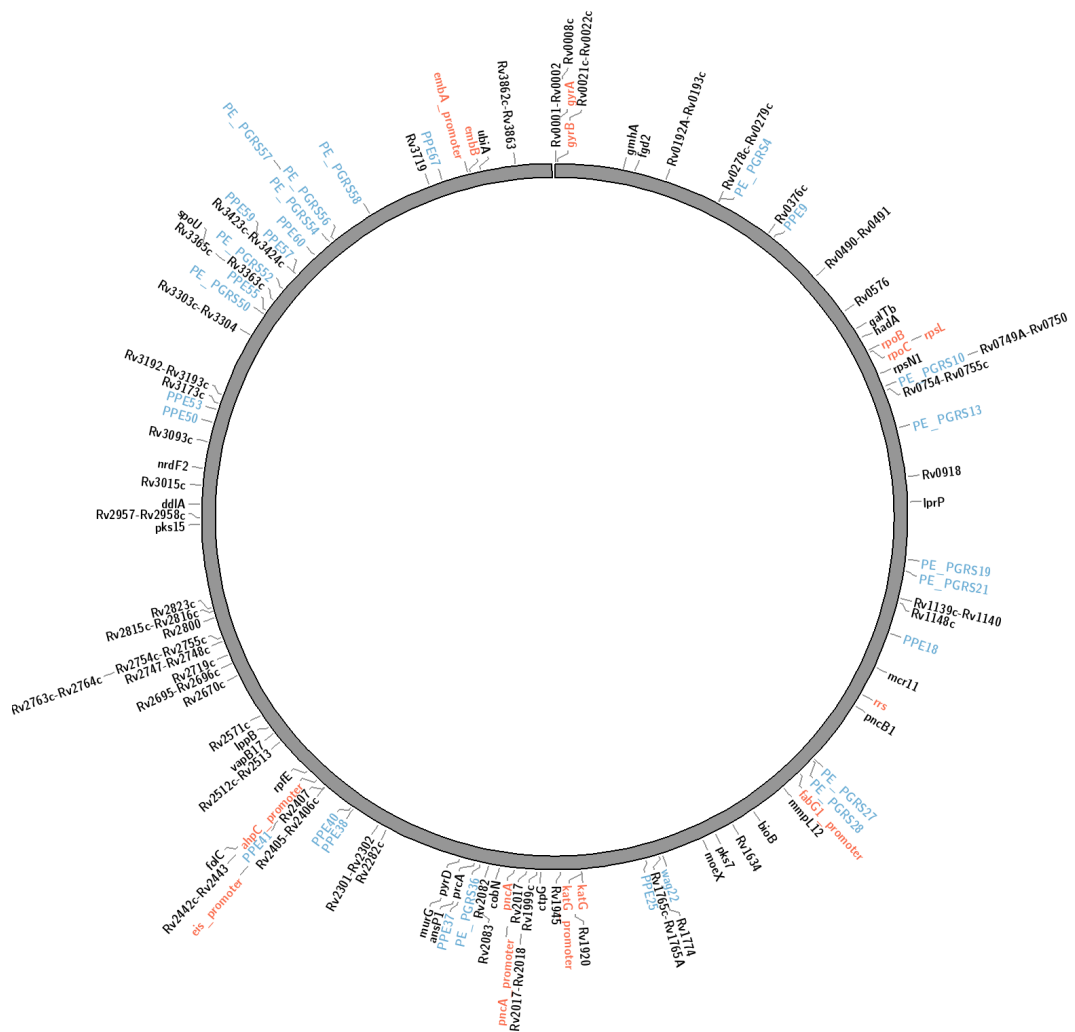


Figure 5.15 *PhyC* results

Loci harbouring more than one *PhyC* SNP hit (p -value < 0.05) are shown, or at least one nsSNP in coding regions. Coding and intergenic regions are placed along the *Mtb* chromosome in this Circos representation and colour-coded in red if previously described to be involved in DR, in blue if they belong to the PE/PPE gene category, or in black if otherwise.

DISCUSSION

Improving our understanding of the relationship between the genotype and the DR phenotype in *Mtb* will aid in the development of more accurate molecular diagnostics for drug-resistant TB. Such insights may be used to overcome or complement the limitations of phenotypic susceptibility testing. The identified associations can shed light on the molecular mechanisms underlying DR and assist in the design of novel antibiotics (Lee *et al.* 2014).

WGS offers the opportunity to capture the genomic diversity of *Mtb* clinical specimens which, coupled with an accurate phenotype characterisation, can be used to dissect the genetic determinants of DR. In this context, it may be challenging to pin down new DR loci because of the diverse genetic background of clinical strains (S. Zhang *et al.* 2013). To overcome this limitation, a large collection of whole-genome sequenced *Mtb* clinical isolates (n=2,765) was employed in this work, covering the four major *Mtb* lineages and progressively resistant isolates from independent studies. It was also ensured that the final dataset had a sizable representation of pan-susceptible samples with diverse genetic backgrounds.

First, the proportion of samples with mutations in known DR candidate genes that could explain the observed phenotypic resistance was calculated. This preliminary analysis showed that a subset of resistant strains lacked mutations in these genes, in line with previous observations (Bhujra *et al.* 2013; Brossier *et al.* 2011; Safi *et al.* 2010), and emphasises the need of improving our understanding of the genetic basis of DR. With this aim in mind, a genome-wide association analysis was conducted to identify new loci closely associated with DR.

Rather than testing each mutation independently for association, genomic variants (SNPs) were aggregated by gene and intergenic region (only nsSNPs in the case of protein coding regions). As a complementary approach, SNPs were also aggregated by operon in order to assess the combined contribution of SNPs in multiple genes (and promoter) likely to be acting within the same biological pathway.

Both locus and operon-based GWAS proved to be successful at identifying well-known DR determinants, which validates the use and implementation of these approaches. In addition to classical DR genes, other classes of genes and operons likely to confer, contribute to or compensate for phenotypic resistance were also identified. The discovery of membrane transporters and efflux pumps is not accidental, as their activity is expected to affect drug transport and therefore drug susceptibility (Black *et al.* 2014; Balganesh *et al.* 2012; Machado *et al.* 2012; Danilchanka *et al.* 2008). Although comprehensively studied *in vitro*, the precise role of these transporters in clinically resistant isolates is controversial (Black *et al.* 2014). The strong association of *Rv0194* with ETH^R provides evidence of a multidrug efflux pump (Danilchanka *et al.* 2008) potentially conferring DR in clinical isolates. The use of inhibitors of efflux pumps has been advocated for preventing the emergence of MDR-TB during treatment (Machado *et al.* 2012).

An intrinsic resistance mechanism to antibiotics results from reduced permeability of the bacterial cell wall. Genes belonging to 'lipid metabolism' and 'cell wall and cell processes' categories are particularly enriched among GWAS and *phyC* hits, including the PE/PPE genes (which encode for a group of secreted proteins), lipoproteins, integral membrane proteins and membrane proteins involved in lipid transport (*mmpL*

genes). These hits are of notable interest given their potential influence in cell wall permeability and fluidity, which in turn determines drug susceptibility (Mdluli *et al.* 1998; Danilchanka *et al.* 2008). On the other hand, it cannot be discounted that these genes are involved in compensatory mechanisms trying to ameliorate the effects of anti-TB drugs on cell wall integrity, as some of the cell wall biosynthetic pathways are indeed targeted by current TB drugs (e.g. INH, ETH or EMB).

The unexpected observation of putative transcriptional regulatory proteins (*Rv1816*, *Rv3736*, *Rv0275c* and *Rv2736c*) associated with DR is an intriguing finding. These transcriptional regulators are found to be intermediately associated with resistance to multiple drugs with the exception of *Rv1816*, being strongly associated with ETH^R. Further work should find out which genes are under their transcriptional control (Galagan *et al.* 2013) as they may underlie more complex mechanisms of DR.

An inherent limitation of this study is the accuracy of phenotypic DSTs, whose results may be unreliable or inconsistent among laboratories. For instance, the great majority of INH^R isolates without known INH^R-conferring mutations (n=90) belong to South Africa (n=56), which may reflect problems in DST determination for this particular population. Still, the finding of well-known DR loci demonstrates the robustness of the analysis to overcome these limitations.

A challenge in DR loci discovery when using *Mtb* clinical isolates is the confounding effect of background resistance. Development of resistance in *Mtb* occurs with the stepwise use of drugs, in response to increasingly resistant TB. Unlike the *phyC* approach, logistic regression adjusted for overlapping resistance managed to remove most of the associations that were likely due to confounding resistance. It should be

noted though, that a drawback of adjusting for co-occurring resistance can be the non-detection of shared DR mechanisms, or removal of signals if the overlap is high. For instance, the association of *rpoC* with RMP^R was lost after adjusting for INH^R (Supplementary Table 10). *gid* and *ethA* genes were intermediately associated with INH and RMP resistance before adjusting for overlapping resistance. However, they were not found to be significantly associated with STR and ETH respectively by neither the GWAS nor the *phyC* approaches. The fact that these two genes are thought to cause intermediate levels of resistance (Morlock & Metchock 2003; Perdigão *et al.* 2013) may explain why there were not detected. DR is frequently treated as a binary variable (resistance or susceptible) although a range of MIC values are indeed measured in clinical isolates. Samples with intermediate levels of resistance, i.e. with MIC values around the cut-off used to determine resistance, may be equally classified as resistant or susceptible. Therefore, the employed association analyses will have limited statistical power to detect loci involved in low or intermediate levels of resistance. The use of quantitative values of DR, i.e. MIC values, and linear regression rather than logistic regression could potentially be more powerful to assess DR phenotype-genotype associations.

Future work could consider approaches that are complementary to GWAS, such as pathway-based methods to assess the combined contribution of multiple genes acting within canonical biological pathways (Eleftherohorinou *et al.* 2009), and logistic regression models including gene epistatic interactions as predictors of DR. Small indels and larger structural variants must also be incorporated in the GWAS. Although systematically neglected in *Mtb* WGS studies, this type of polymorphisms are likely to

have major functional consequences, also in DR (Machado *et al.* 2012), and must therefore be taken into account. Irrespective, any genetic mutations found to be associated with resistance to a specific drug should then be experimentally validated. Functional genetic experiments, like allelic exchange, can establish the casual relationship and elucidate the contribution of each candidate mutation to the DR phenotype (Nebenzahl-Guimaraes *et al.* 2014).

Chapter 6

Discussion and Further Work

6 DISCUSSION AND FURTHER WORK

High throughput sequencing technologies can provide large volumes of WGS data that is complex to analyse and, when combined with meta data, provide epidemiological, antibiotic susceptibility and other clinically-relevant insights. The current state of sequencing technology imposes bioinformatic challenges that need to be addressed in order to allow the transition of bacterial WGS from the research laboratory to a clinical setting (Fricke & Rasko 2013). There is a need for automated bioinformatic workflows that can process the millions of short reads generated by NGS sequencing platforms. While bioinformatic tools for the identification of SNPs and small indels (i.e. shorter than the read length) are now relatively established, the discovery and genotyping of SVs has lagged behind because it is fundamentally more difficult (Alkan *et al.* 2011). As a result, this type of genomic variation has been systematically neglected in *Mtb* WGS studies although it may have greater functional impact due to their larger size. This work describes the implementation of a bioinformatic workflow to effectively extract genetic polymorphisms from WGS data of *Mtb* clinical isolates (Section 2.2.1). In addition to SNPs and small indels, large deletions were discovered by detecting multiple signals in alignment files (i.e. DOC, read-pair and split-read) to increase sensitivity. Although large chromosome rearrangements are rare in MTBC genomes

(Shitikov *et al.* 2014), large insertions (i.e. longer than the read length) are expected to be found in clinical isolates with respect to the H37Rv reference strain. Transposon insertions are a well-known source of genomic variation in MTBC (Reyes *et al.* 2012). The detection of these repetitive elements using short reads is difficult, and they are a known cause of false positive deletions (Alkan *et al.* 2011). Still, a few algorithms have been developed to specifically detect them albeit specific for the human genome (Hormozdiari *et al.* 2010). Novel insertions (i.e. sequences absent in the reference genome) cannot be identified using reference-based mapping approaches. Future work should investigate the accuracy of a whole-genome *de novo* assembly approach for the characterisation of all types of SVs in MTBC strains (Li 2012). Also, such work should attempt to confirm all or a subset of observed variation using another technology, e.g. capillary sequencing. Although, this was not possible in this work because all DNA was used in sequencing, the high coverage achieved and quality control means that the false positive rate is expected to be low.

Some regions of the genome were excluded from analysis. PE and PPE genes are generally excluded from genome analyses because of their high GC content and their repetitive nature (McEvoy *et al.* 2012; Adindla & Guruprasad 2003; Mukhopadhyay & Balaji 2011), which make sequencing and genome assembly difficult. GC-rich regions are under-sampled by Illumina sequencers, resulting in coverage gaps and lower confidence in these regions. In addition, mapping of reads from the PE and PPE genes to the reference genome is prone to errors as the regions are not unique and can result in artefacts, e.g. false positive SNP calls (Nielsen *et al.* 2011). A *de novo* assembly approach would be better at resolving these genes provided that enough depth of

coverage is achieved. A recent study (Bryant, Harris, *et al.* 2013) demonstrated that a large proportion of PE and PPE genes (86%) can be assembled and their genetic variants discovered. Despite accounting for almost 10% of the coding capacity of the genome, the function of PE and PPE genes remains elusive (Mukhopadhyay & Balaji 2011).

The work describes data from first (Sanger capillary) and second-generation (Illumina) sequencing technologies, but a third generation of platforms is under development. Technologies such as Oxford Nanopore (Branton *et al.* 2008; Timp & Mirsaidov 2010) and Single Molecule Real Time (SMRT) PacBio sequencing (Liu *et al.* 2012) promise to deliver longer and more accurate reads at high throughput and reduced costs. The average read length achieved by PacBio sequencers (Pacific Bioscience, California, USA) is 1,300 bp, longer than any current NGS platform (Glenn 2011). As a consequence, algorithms and workflows for sequence analysis will have to be adapted and evolve to accommodate these changes (Satou *et al.* 2014; Boetzer & Pirovano 2014). It can be foreseen that, if these new platforms are increasingly adopted, long-read alignment algorithms (like those employed for capillary sequencing reads) and *de novo* assembly algorithms will become crucial (H. Li & Homer 2010). Longer and more complex SVs, insertion elements and PE/PPE genes will be resolved with better accuracy. Despite the limitations of current Illumina sequencers, 98% of the *Mtb* genome can be mapped uniquely using 75 bp-long reads (compared to 83.1% of the human genome) (Derrien *et al.* 2012). With read lengths of up to 250 bp currently available, an even greater proportion of the *Mtb* genome will be accessible using Illumina sequencers in the near future.

Despite the growing amount of genomic data being generated from *Mtb* clinical isolates, a repository of genetic polymorphisms derived from WGS projects was lacking. In this context, *PolyTB* aims to bring together all existing genomic diversity into an integrated database and make it available for the TB community. Annotation of DR mutations (Section 4.2.1) and strain-specific SNPs (Section 3.2.4) will further enhance the usefulness of *PolyTB*. A database like *PolyTB* should also allow users to upload their own sequenced samples. Genetic variants could be annotated, including DR and strain-specific mutations, and compared to those already present in the database. Functionality enabling the phylogenetic positioning of uploaded samples, i.e. reporting the genetically closest sample in the database, would be a potentially useful addition to the tool.

The high global burden of TB requires new control insights from the increasing number of *Mtb* WGS studies. Knowledge of the genetic diversity across populations, among other factors, will assist in the understanding of *Mtb* biology, required to develop new drugs and novel vaccines. Strain-specific genomic diversity in MTBC is an important factor in pathogenesis that may affect virulence (Nahid *et al.* 2010; Thwaites *et al.* 2008; Caws *et al.* 2008), transmissibility (Kato-Maeda & Kim 2010), host response (López *et al.* 2003) and emergence of DR (Ford *et al.* 2013). Several systems have been proposed to classify MTBC strains into distinct lineages and families, using molecular genotypes (Kamerbeek *et al.* 1997; Supply *et al.* 2001), regions of difference (Gagneux *et al.* 2006) and SNPs. However, classical genotyping lacks robust phylogenetic markers (Comas *et al.* 2009; Roetzer *et al.* 2013) and alternative classification systems based on SNPs lack resolution (Stucki *et al.* 2012) or do not capture known circulating strain-

types (Homolka *et al.* 2012). The SNP barcode developed in this work is the first to cover all main lineages, including the recently discovered lineage 7 (Firdessa *et al.* 2013), and classifies a greater number of sub-lineages than current alternatives. It also outperforms the other SNP systems. One of the latest published SNP sets (Homolka *et al.* 2012) was employed in recent studies to type *Mtb* strains after undergoing WGS (Casali *et al.* 2014; Chernyaeva *et al.* 2014). Still, 36% of isolates could not be classified (Casali *et al.* 2014), particularly those from the Euro-American lineage, for which the global diversity is not fully captured by this scheme. All samples in that study could be unambiguously classified using the SNP barcode in this work (Section 3.3.4). Furthermore, the presented classification scheme is fully compatible with the gold-standard RD system and is comparable with spoligotypes (Table 3.3). The SNP barcode developed has the potential to be applied in epidemiological and surveillance settings, but this would require the markers to be incorporated into genotyping platforms, such as multiplex ligation-dependent probe amplification (MLPA) assays (Bergval *et al.* 2012; Sengstake *et al.* 2014), multiplexed oligonucleotides ligation PCR (MOL-PCR) (Deshpande *et al.* 2010), or TaqMan real-time PCR (Stucki *et al.* 2012). Since these methods differ in their technical requirements, different SNP sets may be required depending on the particular platform (Kim & Misra 2007). Still, given the redundancy of phylogenetically informative SNPs discovered in this work (for both lineages and sub-lineages), different phylogenetically equivalent SNP sets could be chosen for each SNP-typing platform.

The proposed barcode could also be used to inform genotype-phenotype association studies such as those investigating host-pathogen interactions during TB infection, wherein strain-type is likely to be an important factor in pathogenesis. It is known that population structure can lead to false positive associations when considering phenotypes like drug resistance, and knowledge of strain-specific markers can minimise the spurious findings. In the context of clinical trials, the barcode could be applied to evaluate new TB vaccines whose protective efficacy may vary by the genotype of the infecting strain (López *et al.* 2003).

Despite the limitations of current genotyping techniques (i.e. spoligotyping, MIRU-VNTR and RFLP) compared to SNP-typing, they are still broadly employed in reference and research laboratories. There are available databases with hundreds of MTBC isolates typed using classical genotyping (Demay *et al.* 2012), and their strain nomenclature widely employed in the literature (e.g. LAM, Haarlem or Beijing). *SpolPred* has proved to be a useful tool to predict spoligotype from WGS data. In addition to being used throughout this work, it has been employed by others. For example, in epidemiological studies where experimental genotypic data may be present (Guerra-Assunção *et al.* 2014), thus a source of confirmation, or absent (Liu *et al.* 2014; Rashdi & Jadhav 2014) where a strain-type needs to be identified. Future work may consider predicting MIRU-VNTR patterns from WGS data. Their *in silico* determination from short sequencing reads is expected to be computationally challenging as the current read lengths (50-100 bp) may not span multiple VNTR repeats. Determination of RFLP patterns would rely on an accurate reconstruction of the whole genome, followed by *in silico* 'digestion' of the WGS with the restriction

enzyme and estimation of the bands length. Irrespective, *in silico* typing approaches for *Mtb* could be tested and validated as datasets with both the experimentally determined genotype and WGS are available (Pérez-Lago *et al.* 2014; Bryant, Harris, *et al.* 2013; Guerra-Assunção *et al.* 2014). Similar approaches, particularly aimed at predicting multi-locus sequence types (MLST) from short reads, have also been developed for other bacterial genomes (Inouye *et al.* 2012).

Combining strain identification with drug susceptibility determination will further enhance the usefulness of the barcode. Genetic profiles of resistance are valuable clinical information to design tailored drug regimens. Knowledge of the genetic background of circulating strains can provide insights into DR emergence and spread. In that respect, the identification of similar strains in more than one patient sharing DR mutations is likely to reflect transmission of DR between them, and allow early interventions to avoid onward transmission. In addition, associations of particular strain types and/or DR profiles with poor treatment outcomes or mortality could also be determined. A rapid genetic test incorporating strain-specific and DR mutations (using WGS or another genotyping platform) would be beneficial for therapeutic selection, clinical management of patients and infection control measures.

Current phenotypic DSTs require isolation from sputum and culture of *Mtb* followed by exposure to anti-TB drugs, a process that may take weeks or months and requires high levels of microbiological safety. Phenotypic DSTs are technically complex, which can lead to variable reliability among laboratories. Rapid molecular assays are now available for some key drugs. They examine a limited number of loci and normally the

most frequent DR mutations only, which may result in low sensitivity, especially in certain geographical regions where the most common DR mutations are less prevalent. This observation is the case for INH detection using *MDRTBplus* (Jin *et al.* 2012; Akpaka *et al.* 2008). Furthermore, they do not differentiate silent mutations from those conferring resistance (Alonso *et al.* 2011; Jin *et al.* 2013; Aubry *et al.* 2014), which will lead to false positives.

Development of new and more accurate genetic DSTs will rely on the understanding of the relationships between the genotype and DR phenotype. This means that phenotypic DSTs are not replaceable for drugs without fully characterised genetic mechanisms of resistance and for recently licensed drugs like bedaquiline, for which DR mutations in clinical samples have not been established. Nevertheless, WGS can be used to rapidly identify resistance when mutations known to be associated with resistance are detected. In this regard, the presented library of DR mutations (Section 4.2.1) is the most complete and updated database of its kind. The DR markers in this list can be then used to diagnose DR-TB from WGS data as shown (4.3.1) or be incorporated in alternative genotyping platforms for the same purpose.

In addition to its potential as a diagnostic tool, WGS opens the possibility of deciphering novel mechanisms of antibiotic resistance. The complementary GWAS and selection detection methodology described in this work to assess genotype-phenotype association identified well-known drug targets and drug-converting enzymes. It appears that potential compensatory mechanisms (e.g. *rpoC*) may be better detected using the phylogenetic-based selection metrics. This work, as well as (Farhat *et al.*

2013; H. Zhang *et al.* 2013; Safi *et al.* 2013), identifies new resistance-associated loci and expands our understanding of the genetic diversity underlying DR (Warner & Mizrahi 2013). It further highlights that the genetic basis of DR phenotypes can be more complex than previously anticipated, and the phenotype may develop through the contribution of multiple loci, resulting in a range of drug susceptibilities (Safi *et al.* 2013). In light of these observations, the binary classification of *Mtb* as either susceptible or resistance, rather than consideration as a continuous spectrum, is an oversimplification as it does not account for different levels of resistance present in clinical strains, and complicates DR association analyses. Future work should consider using quantitative values of DR, such as MIC values. A linear regression rather than logistic regression would be used to assess DR associations between mutations and loci. Additional analysis could also take association p-values and look for biological pathways that may be over-represented. However, such analysis is dependent on well-characterised pathway information in *Mtb*.

WGS offers the opportunity to capture the natural genetic variation in clinical specimens and identify the genetic mutations associated with a specific trait, which can be then experimentally validated. Allelic exchange experiments can follow GWAS to establish the casual relationship and elucidate the contribution of each mutation to the DR phenotype (Nebenzahl-Guimaraes *et al.* 2014). This approach is more likely to succeed than performing directed evolution experiments to characterise *in vitro* acquired DR mutations, which may not be ever observed in clinical samples.

The combination of genomics with other 'omic' approaches, such as proteomics or metabolomics, opens the door to the study of more complex phenotypes, including host-pathogen interactions. These interactions are likely to be complex, between genes that are not apparent from the genome sequence alone. Transcriptional profiling of wild-type and mutant strains – initially performed using microarrays (Butcher 2004) and more recently with RNA-seq - has been extensively applied to study the adaptive responses of *Mtb*. In particular, adaptations under certain experimental conditions, such as antibiotic stress (Waddell & Butcher 2010), uptake by macrophages (Monahan *et al.* 2001) or reduction of oxygen and other nutrients (Rodríguez & Hernández 2014). RNA-seq has facilitated the mapping of transcriptional start sites (TSSs) (Cortes *et al.* 2013) and identification of non-coding RNAs in the *Mtb* genome (Miotto *et al.* 2012). A recent study used ChIP-seq (Chromatin Immunoprecipitation and Sequencing) in *Mtb* to map transcription factor binding sites across the genome for the first time (Galagan *et al.* 2013), resulting in a far more complex and interconnected regulatory network than previously anticipated. The combination of RNA-seq with ChIP-seq (Uplekar *et al.* 2013) and proteomics (Cortes *et al.* 2013) can shed light on more complex mechanisms of transcriptional and post-transcription regulation of gene expression. Systems biology approaches can provide a deeper understanding of pathogen-host interactions, especially those involved in latency, reactivation and immune response - needed to develop better TB drugs and more effective vaccines.

Rapid, low-cost genome sequencing is having a big impact on molecular epidemiology, evolution and diagnosis of bacterial infections, enabling researchers and clinicians to

gain insights at the patient, community and global levels. The potential application of WGS to diagnose antibiotic resistance and other clinically-relevant phenotypes is an area of active research. Given the cost decline witnessed in recent years (Sboner *et al.* 2011) and advances in sequencing directly from clinical samples (Köser *et al.* 2013), it is foreseen that WGS will be eventually a technology of choice in clinical settings. In this new paradigm, the presented work will facilitate the transition to and applications of WGS in clinical settings as an important tool for TB control. These possibilities have special significance in regions of the world, such as southern Africa, where TB continues to claim thousands of lives every year.

REFERENCES

- Abadia, E. *et al.*, 2010. Resolving lineage assignment on Mycobacterium tuberculosis clinical isolates classified by spoligotyping with a new high-throughput 3R SNPs based method. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 10(7), pp.1066–74. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/20624486> [Accessed June 8, 2014].
- Abadia, E. *et al.*, 2011. The use of microbead-based spoligotyping for Mycobacterium tuberculosis complex to evaluate the quality of the conventional method: providing guidelines for Quality Assurance when working on membranes. *BMC infectious diseases*, 11(1), p.110. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3107175&tool=pmcentrez&rendertype=abstract> [Accessed January 9, 2012].
- Abubakar, I. *et al.*, 2013. Drug-resistant tuberculosis: time for visionary political leadership. *Lancet Infect Dis*, 13(6), pp.529–39. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23531391> [Accessed August 6, 2013].
- Abyzov, A. *et al.*, 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.*, 21(6), pp.974–84. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3106330&tool=pmcentrez&rendertype=abstract> [Accessed July 12, 2012].
- Abyzov, A. & Gerstein, M., 2011. AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics*, 27(5), pp.595–603. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3042181&tool=pmcentrez&rendertype=abstract> [Accessed July 14, 2012].
- Adindla, S. & Guruprasad, L., 2003. Sequence analysis corresponding to the PPE and PE proteins in Mycobacterium tuberculosis and other genomes. *Journal of biosciences*, 28(2), pp.169–79. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/12711809> [Accessed January 18, 2014].
- Ajbani, K. *et al.*, 2012. Evaluation of genotype MTBDRsl assay to detect drug resistance associated with fluoroquinolones, aminoglycosides and ethambutol on clinical sediments. *PloS one*, 7(11), p.e49433. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3499545&tool=pmcentrez&rendertype=abstract> [Accessed October 30, 2013].
- Akpaka, P.E. *et al.*, 2008. Evaluation of methods for rapid detection of resistance to isoniazid and rifampin in Mycobacterium tuberculosis isolates collected in the Caribbean. *Journal of clinical microbiology*, 46(10), pp.3426–8. Available at: <http://jcm.asm.org/content/46/10/3426.full> [Accessed July 15, 2014].
- Alam, M.T. *et al.*, 2014. Dissecting vancomycin intermediate resistance in Staphylococcus aureus using genome-wide association. *Genome biology and evolution*. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24787619> [Accessed May 6, 2014].
- Alix, E., Godreuil, S. & Blanc-Potard, A.-B., 2006. Identification of a Haarlem genotype-specific single nucleotide polymorphism in the mgtC virulence gene of Mycobacterium tuberculosis. *Journal of clinical microbiology*, 44(6), pp.2093–8. Available at:

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1489410&tool=pmcentrez&rendertype=abstract> [Accessed March 26, 2014].

- Alkan, C., Coe, B.P. & Eichler, E.E., 2011. Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5), pp.363–76. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/21358748> [Accessed October 26, 2012].
- Alonso, M. *et al.*, 2011. Isolation of Mycobacterium tuberculosis strains with a silent mutation in rpoB leading to potential misassignment of resistance category. *Journal of clinical microbiology*, 49(7), pp.2688–90. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3147855&tool=pmcentrez&rendertype=abstract> [Accessed April 29, 2014].
- Altman, D.G., 1990. *Practical Statistics for Medical Research*, CRC Press. Available at: http://books.google.co.uk/books/about/Practical_Statistics_for_Medical_Research.html?id=v-walRnRxWQC&pgis=1 [Accessed August 13, 2014].
- Ando, H. *et al.*, 2011. Downregulation of katG expression is associated with isoniazid resistance in Mycobacterium tuberculosis. *Molecular microbiology*, 79(6), pp.1615–28. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/21244531> [Accessed September 28, 2013].
- Ando, H. *et al.*, 2010. Pyrazinamide resistance in multidrug-resistant Mycobacterium tuberculosis isolates in Japan. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 16(8), pp.1164–8. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/19832709>.
- Andries, K. *et al.*, 2005. A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis. *Science (New York, N.Y.)*, 307(5707), pp.223–7. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/15591164> [Accessed July 9, 2014].
- Angeby, K. a *et al.*, 2010. Wild-type MIC distributions of four fluoroquinolones active against Mycobacterium tuberculosis in relation to current critical concentrations and available pharmacokinetic and pharmacodynamic data. *The Journal of antimicrobial chemotherapy*, 65(5), pp.946–52. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/20332195> [Accessed April 11, 2014].
- Anon, 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145), pp.661–78. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2719288&tool=pmcentrez&rendertype=abstract> [Accessed July 9, 2014].
- Aubry, A. *et al.*, 2014. First Evaluation of Drug-Resistant Mycobacterium tuberculosis Clinical Isolates from Congo Revealed Misdetection of Fluoroquinolone Resistance by Line Probe Assay Due to a Double Substitution T80A-A90G in GyrA. *PloS one*, 9(4), p.e95083. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3990612&tool=pmcentrez&rendertype=abstract> [Accessed May 17, 2014].
- Balganesh, M. *et al.*, 2012. Efflux pumps of Mycobacterium tuberculosis play a significant role in antituberculosis activity of potential drug candidates. *Antimicrobial agents and chemotherapy*, 56(5), pp.2643–51. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3346595&tool=pmcentrez&rendertype=abstract> [Accessed June 5, 2014].

- Banerjee, A. *et al.*, 1994. inhA, a gene encoding a target for isoniazid and ethionamide in *Mycobacterium tuberculosis*. *Science (New York, N.Y.)*, 263(5144), pp.227–30. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/8284673> [Accessed May 18, 2014].
- Barry, C.E. *et al.*, 2009. The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. *Nature reviews. Microbiology*, 7(12), pp.845–55. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/19855401> [Accessed May 25, 2014].
- Baulard, A.R. *et al.*, 2000. Activation of the pro-drug ethionamide is regulated in mycobacteria. *The Journal of biological chemistry*, 275(36), pp.28326–31. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/10869356> [Accessed May 12, 2014].
- Bentley, S.D. *et al.*, 2012. The genome of *Mycobacterium africanum* West African 2 reveals a lineage-specific locus and genome erosion common to the *M. tuberculosis* complex. *PLoS neglected tropical diseases*, 6(2), p.e1552. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3289620&tool=pmcentrez&rendertype=abstract> [Accessed July 12, 2012].
- Berg, J.S., Khoury, M.J. & Evans, J.P., 2011. Deploying whole genome sequencing in clinical practice and public health: meeting the challenge one bin at a time. *Genetics in medicine : official journal of the American College of Medical Genetics*, 13(6), pp.499–504. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/21558861> [Accessed May 13, 2014].
- Bergval, I. *et al.*, 2012. Combined species identification, genotyping, and drug resistance detection of *Mycobacterium tuberculosis* cultures by MLPA on a bead-based array. *PLoS One*, 7(8), p.e43240. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3423362&tool=pmcentrez&rendertype=abstract> [Accessed November 1, 2012].
- Bhaju, S. *et al.*, 2013. *Mycobacterium tuberculosis* isolates from Rio de Janeiro reveal unusually low correlation between pyrazinamide resistance and mutations in the pncA gene. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 19C, pp.1–6. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23770140> [Accessed September 28, 2013].
- Black, P. a *et al.*, 2014. Energy metabolism and drug efflux in *Mycobacterium tuberculosis*. *Antimicrobial agents and chemotherapy*, 58(5), pp.2491–503. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24614376> [Accessed May 26, 2014].
- Blouin, Y. *et al.*, 2012. Significance of the Identification in the Horn of Africa of an Exceptionally Deep Branching *Mycobacterium tuberculosis* Clade. *PLoS one*, 7(12), p.e52841. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531362&tool=pmcentrez&rendertype=abstract> [Accessed February 3, 2013].
- Boetzer, M. & Pirovano, W., 2014. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC bioinformatics*, 15, p.211. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4076250&tool=pmcentrez&rendertype=abstract> [Accessed July 11, 2014].
- Boonaiam, S. *et al.*, 2010. Genotypic analysis of genes associated with isoniazid and ethionamide resistance in MDR-TB isolates from Thailand. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 16(4), pp.396–9. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/19486070>.

- Branton, D. *et al.*, 2008. The potential and challenges of nanopore sequencing. *Nature biotechnology*, 26(10), pp.1146–53. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2683588&tool=pmcentrez&rendertype=abstract> [Accessed October 26, 2012].
- Brennan, P.J., 2003. Structure, function, and biogenesis of the cell wall of *Mycobacterium tuberculosis*. *Tuberculosis (Edinburgh, Scotland)*, 83(1-3), pp.91–7. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/12758196> [Accessed June 2, 2014].
- Brosch, R. *et al.*, 2007. Genome plasticity of BCG and impact on vaccine efficacy. *Proceedings of the National Academy of Sciences of the United States of America*, 104(13), pp.5596–601. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1838518&tool=pmcentrez&rendertype=abstract>.
- Brossier, F. *et al.*, 2011. Molecular investigation of resistance to the antituberculous drug ethionamide in multidrug-resistant clinical isolates of *Mycobacterium tuberculosis*. *Antimicrobial agents and chemotherapy*, 55(1), pp.355–60. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3019671&tool=pmcentrez&rendertype=abstract> [Accessed October 4, 2013].
- Bryant, J.M., Schürch, A.C., *et al.*, 2013. Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. *BMC Infect. Dis.*, 13(1), p.110. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23446317>.
- Bryant, J.M., Harris, S.R., *et al.*, 2013. Whole-genome sequencing to establish relapse or re-infection with *Mycobacterium tuberculosis*: a retrospective observational study. *The Lancet Respiratory Medicine*, 1(10), pp.786–792. Available at: <http://linkinghub.elsevier.com/retrieve/pii/S2213260013702315> [Accessed January 16, 2014].
- Butcher, P.D., 2004. Microarrays for *Mycobacterium tuberculosis*. *Tuberculosis (Edinburgh, Scotland)*, 84(3-4), pp.131–7. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/15207482> [Accessed August 12, 2014].
- Camacho, C. *et al.*, 2009. BLAST+: architecture and applications. *BMC bioinformatics*, 10, p.421. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2803857&tool=pmcentrez&rendertype=abstract> [Accessed July 9, 2011].
- Casali, N. *et al.*, 2014. Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nature genetics*, 46(3), pp.279–86. Available at: <http://dx.doi.org/10.1038/ng.2878> [Accessed March 19, 2014].
- Casali, N. & Nikolayevskyy, V., 2012. Microevolution of extensively drug-resistant tuberculosis in Russia. *Genome Res.*, 22(4), pp.735–745. Available at: <http://genome.cshlp.org/content/22/4/735.short> [Accessed November 1, 2012].
- Caws, M. *et al.*, 2008. The influence of host and bacterial genotype on the development of disseminated disease with *Mycobacterium tuberculosis*. *PLoS pathogens*, 4(3), p.e1000034. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2268004&tool=pmcentrez&rendertype=abstract> [Accessed December 16, 2013].
- Chackerian, A.A. *et al.*, 2002. Dissemination of *Mycobacterium tuberculosis* is influenced by host factors and precedes the initiation of T-cell immunity. *Infection and immunity*, 70(8),

- pp.4501–9. Available at:
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=128141&tool=pmcentrez&rendertype=abstract> [Accessed August 3, 2014].
- Chen, K. *et al.*, 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, 6(9), pp.677–81. Available at:
<http://www.ncbi.nlm.nih.gov/pubmed/19668202> [Accessed June 14, 2011].
- Chernyaeva, E.N. *et al.*, 2014. Genome-wide Mycobacterium tuberculosis variation (GMTV) database: a new tool for integrating sequence variations and epidemiology. *BMC genomics*, 15(1), p.308. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24767249> [Accessed April 29, 2014].
- Clark, T.G. *et al.*, 2013. Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing. *PLoS one*, 8(12), p.e83012. Available at:
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3859632&tool=pmcentrez&rendertype=abstract> [Accessed January 9, 2014].
- Cock, P.J.A. *et al.*, 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6), pp.1767–1771. Available at:
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2847217&tool=pmcentrez&rendertype=abstract>.
- Cole, S. *et al.*, 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, 393(6685), pp.357–44. Available at:
<http://www.nature.com/nature/journal/v393/n6685/abs/393537a0.html> [Accessed July 8, 2012].
- Coll, F. *et al.*, 2014. PolyTB: A genomic variation map for Mycobacterium tuberculosis. *Tuberculosis (Edinburgh, Scotland)*, pp.1–9. Available at:
<http://www.ncbi.nlm.nih.gov/pubmed/24637013> [Accessed March 19, 2014].
- Coll, F. *et al.*, 2012. SpolPred: rapid and accurate prediction of Mycobacterium tuberculosis spoligotypes from short genomic sequences. *Bioinformatics*, 28(22), pp.2991–3. Available at:
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3496340&tool=pmcentrez&rendertype=abstract> [Accessed January 29, 2013].
- Comas, I. *et al.*, 2009. Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies. *PLoS one*, 4(11), p.e7815. Available at:
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2772813&tool=pmcentrez&rendertype=abstract> [Accessed March 28, 2014].
- Comas, I. *et al.*, 2010. Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. *Nature genetics*, 42(6), pp.498–503. Available at:
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2883744&tool=pmcentrez&rendertype=abstract> [Accessed November 11, 2013].
- Comas, I. *et al.*, 2013. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nat. Genet.*, 45(10), pp.1176–82. Available at:
<http://www.nature.com/doifinder/10.1038/ng.2744> [Accessed September 3, 2013].

- Comas, I. & Gagneux, S., 2009. The past and future of tuberculosis research. *PLoS pathogens*, 5(10), p.e1000600. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2745564&tool=pmcentrez&rendertype=abstract> [Accessed March 10, 2012].
- Cooper, A.M., Mayer-Barber, K.D. & Sher, A., 2011. Role of innate cytokines in mycobacterial infection. *Mucosal immunology*, 4(3), pp.252–60. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3294290&tool=pmcentrez&rendertype=abstract> [Accessed July 28, 2014].
- Copin, R., Coscollá, M. & Seiffert, S., 2014. Sequence Diversity in the *pe_pgrs* Genes of *Mycobacterium tuberculosis* Is Independent of Human T Cell Recognition. *mBio*. Available at: <http://mbio.asm.org/content/5/1/e00960-13.short> [Accessed January 26, 2014].
- Cortes, T. *et al.*, 2013. Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in *Mycobacterium tuberculosis*. *Cell reports*, 5(4), pp.1121–31. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24268774> [Accessed January 17, 2014].
- Danilchanka, O., Mailaender, C. & Niederweis, M., 2008. Identification of a novel multidrug efflux pump of *Mycobacterium tuberculosis*. *Antimicrobial agents and chemotherapy*, 52(7), pp.2503–11. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2443884&tool=pmcentrez&rendertype=abstract> [Accessed June 6, 2014].
- Davis, J.L. *et al.*, 2013. Diagnostic accuracy of same-day microscopy versus standard microscopy for pulmonary tuberculosis: a systematic review and meta-analysis. *The Lancet infectious diseases*, 13(2), pp.147–54. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3836432&tool=pmcentrez&rendertype=abstract> [Accessed July 31, 2014].
- DeBarber, a E. *et al.*, 2000. Ethionamide activation and sensitivity in multidrug-resistant *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences of the United States of America*, 97(17), pp.9677–82. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=16924&tool=pmcentrez&rendertype=abstract>.
- Demay, C. *et al.*, 2012. SITVITWEB - A publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. *Infect. Genet. Evol.*, 12(4), pp.755–766. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/22365971> [Accessed March 25, 2012].
- Derrien, T. *et al.*, 2012. Fast computation and applications of genome mappability. *PloS one*, 7(1), p.e30377. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3261895&tool=pmcentrez&rendertype=abstract> [Accessed October 30, 2012].
- Deshpande, A. *et al.*, 2010. A rapid multiplex assay for nucleic acid-based diagnostics. *Journal of microbiological methods*, 80(2), pp.155–63. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/20006656> [Accessed July 14, 2014].
- Van Deun, A. *et al.*, 2013. Rifampin drug resistance tests for tuberculosis: challenging the gold standard. *Journal of clinical microbiology*, 51(8), pp.2633–40. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3719626&tool=pmcentrez&rendertype=abstract> [Accessed May 15, 2014].

- Dheda, K. *et al.*, 2014. Global control of tuberculosis: from extensively drug-resistant to untreatable tuberculosis. *The lancet. Respiratory medicine*, 2(4), pp.321–338. Available at: [http://www.thelancet.com/journals/a/article/PIIS2213-2600\(14\)70031-1/fulltext](http://www.thelancet.com/journals/a/article/PIIS2213-2600(14)70031-1/fulltext) [Accessed May 6, 2014].
- Eleftherohorinou, H. *et al.*, 2009. Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases. *PloS one*, 4(11), p.e8068. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2778995&tool=pmcentrez&rendertype=abstract> [Accessed November 1, 2012].
- Van Embden, J.D. *et al.*, 2000. Genetic variation and evolutionary origin of the direct repeat locus of Mycobacterium tuberculosis complex bacteria. *Journal of bacteriology*, 182(9), pp.2393–401. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=111299&tool=pmcentrez&rendertype=abstract>.
- Engohang-Ndong, J. *et al.*, 2003. EthR, a repressor of the TetR/CamR family implicated in ethionamide resistance in mycobacteria, octamerizes cooperatively on its operator. *Molecular Microbiology*, 51(1), pp.175–188. Available at: <http://doi.wiley.com/10.1046/j.1365-2958.2003.03809.x> [Accessed October 4, 2013].
- Engström, A. *et al.*, 2012. Detection of first- and second-line drug resistance in Mycobacterium tuberculosis clinical isolates by pyrosequencing. *Journal of clinical microbiology*, 50(6), pp.2026–33. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3372151&tool=pmcentrez&rendertype=abstract> [Accessed September 29, 2013].
- Farhat, M.R. *et al.*, 2013. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. *Nature genetics*, 45(10), pp.1183–1189. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23995135> [Accessed September 17, 2013].
- Feasey, N.A. *et al.*, 2011. Moxifloxacin and pyrazinamide susceptibility testing in a complex case of multidrug-resistant tuberculosis. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease*, 15(3), pp.417–20. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/21333115> [Accessed August 12, 2014].
- Felsenstein, J., 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, 5, pp.164–166.
- Feuerriegel, S. *et al.*, 2010. Thr202Ala in thyA is a marker for the Latin American Mediterranean lineage of the Mycobacterium tuberculosis complex rather than para-aminosalicylic acid resistance. *Antimicrobial agents and chemotherapy*, 54(11), pp.4794–8. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2976163&tool=pmcentrez&rendertype=abstract> [Accessed March 26, 2014].
- Feuerriegel, S., Köser, C.U. & Niemann, S., 2014. Phylogenetic polymorphisms in antibiotic resistance genes of the Mycobacterium tuberculosis complex. *The Journal of antimicrobial chemotherapy*. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24458512> [Accessed February 26, 2014].
- Filliol, I., 2000. Detection of a previously unamplified spacer within the DR locus of Mycobacterium tuberculosis: epidemiological implications. *Journal of clinical microbiology*, 38(3), pp.1231–4. Available at:

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=86384&tool=pmcentrez&rendertype=abstract> [Accessed August 15, 2012].

- Filliol, I. *et al.*, 2006. Global Phylogeny of Mycobacterium tuberculosis Based on Single Nucleotide Polymorphism (SNP) Analysis : Insights into Tuberculosis Evolution , Phylogenetic Accuracy of Other DNA Fingerprinting Systems , and Recommendations for a Minimal Standard SNP Set. *J. Bacteriol.*, 188(2), pp.759–772.
- Fine, P.E., 1995. Variation in protection by BCG: implications of and for heterologous immunity. *Lancet*, 346(8986), pp.1339–45. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/7475776> [Accessed August 13, 2014].
- Firdessa, R., Berg, S. & Hailu, E., 2013. Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia. *Emerging infectious ...*, 19(3), pp.460–463. Available at: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3647644/> [Accessed March 31, 2014].
- Flandrois, J.-P., Lina, G. & Dumitrescu, O., 2014. MUBII-TB-DB: a database of mutations associated with antibiotic resistance in Mycobacterium tuberculosis. *BMC bioinformatics*, 15(1), p.107. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24731071> [Accessed April 30, 2014].
- Flynn, J.L., Chan, J. & Lin, P.L., 2011. Macrophages and control of granulomatous inflammation in tuberculosis. *Mucosal immunology*, 4(3), pp.271–8. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3311958&tool=pmcentrez&rendertype=abstract> [Accessed August 3, 2014].
- Ford, C. *et al.*, 2012. Mycobacterium tuberculosis - Heterogeneity revealed through whole genome sequencing. *Tuberculosis (Edinb)*, 92(3), pp.194–201. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/22218163> [Accessed January 9, 2012].
- Ford, C.B. *et al.*, 2013. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nature genetics*, 45(7), pp.784–90. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23749189> [Accessed November 14, 2013].
- Fricke, W.F. & Rasko, D. a, 2013. Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nature reviews. Genetics*, 15(1), pp.49–55. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24281148> [Accessed December 12, 2013].
- Gagneux, S. *et al.*, 2006. Variable host-pathogen compatibility in Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences of the United States of America*, 103(8), pp.2869–73. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1413851&tool=pmcentrez&rendertype=abstract>.
- Gagneux, S. & Small, P.M., 2007. Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. *The Lancet infectious diseases*, 7(5), pp.328–337. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/17448936> [Accessed October 24, 2011].
- Galagan, J.E., 2014. Genomic insights into tuberculosis. *Nature reviews. Genetics*, 15(5), pp.307–20. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24662221> [Accessed May 28, 2014].

- Galagan, J.E. *et al.*, 2013. The Mycobacterium tuberculosis regulatory network and hypoxia. *Nature*, 499(7457), pp.178–83. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4087036&tool=pmcentrez&rendertype=abstract> [Accessed July 16, 2014].
- Garcia-Betancur, J.C. *et al.*, 2011. Alignment of multiple complete genomes suggests that gene rearrangements may contribute towards the speciation of Mycobacteria. *Infect. Genet. Evol.*, 12(4), pp.819–26. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/22008279> [Accessed October 24, 2011].
- García-Sierra, N. *et al.*, 2011. Pyrosequencing for rapid molecular detection of rifampin and isoniazid resistance in Mycobacterium tuberculosis strains and clinical specimens. *Journal of clinical microbiology*, 49(10), pp.3683–6. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3187309&tool=pmcentrez&rendertype=abstract> [Accessed October 30, 2013].
- Gardy, J.L. *et al.*, 2011. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N. Engl. J. Med.*, 364(8), pp.730–9. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/21345102>.
- Georghiou, S.B. *et al.*, 2012. Evaluation of genetic mutations associated with Mycobacterium tuberculosis resistance to amikacin, kanamycin and capreomycin: a systematic review. *PloS one*, 7(3), p.e33275. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3315572&tool=pmcentrez&rendertype=abstract> [Accessed September 19, 2013].
- Glenn, T.C., 2011. Field guide to next-generation DNA sequencers. *Molecular ecology resources*, 11(5), pp.759–69. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/21592312> [Accessed March 8, 2012].
- Gomes, L.H.F. *et al.*, 2011. Genome sequence of Mycobacterium bovis BCG Moreau, the Brazilian vaccine strain against tuberculosis. *Journal of bacteriology*, 193(19), pp.5600–1. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3187452&tool=pmcentrez&rendertype=abstract> [Accessed September 21, 2012].
- Guerra-Assunção, J. *et al.*, 2014. Large scale population-based whole genome sequencing of Mycobacterium tuberculosis provides insights into transmission in a high prevalence area. *In preparation*.
- Harismendy, O. *et al.*, 2013. Evaluation of ultra-deep targeted sequencing for personalized breast cancer care. *Breast cancer research : BCR*, 15(6), p.R115. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3978701&tool=pmcentrez&rendertype=abstract> [Accessed May 17, 2014].
- Hartkoorn, R.C., Uplekar, S. & Cole, S.T., 2014. Cross-Resistance between Clofazimine and Bedaquiline through Upregulation of MmpL5 in Mycobacterium tuberculosis. *Antimicrobial agents and chemotherapy*, 58(5), pp.2979–81. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24590481> [Accessed May 17, 2014].
- Helb, D. *et al.*, 2010. Rapid detection of Mycobacterium tuberculosis and rifampin resistance by use of on-demand, near-patient technology. *Journal of clinical microbiology*, 48(1), pp.229–37. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2812290&tool=pmcentrez&rendertype=abstract> [Accessed October 30, 2013].

- Homolka, S. *et al.*, 2012. High resolution discrimination of clinical Mycobacterium tuberculosis complex strains based on single nucleotide polymorphisms. *PloS one*, 7(7), p.e39855. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3388094&tool=pmcentrez&rendertype=abstract> [Accessed March 28, 2014].
- Homolka, S. *et al.*, 2012. High Resolution Discrimination of Clinical Mycobacterium tuberculosis Complex Strains Based on Single Nucleotide Polymorphisms. *PloS One*, 7(7), p.e39855.
- Hormozdiari, F. *et al.*, 2010. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics (Oxford, England)*, 26(12), pp.i350–7. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2881400&tool=pmcentrez&rendertype=abstract> [Accessed July 5, 2011].
- Ilna, E.N. *et al.*, 2013. Comparative genomic analysis of Mycobacterium tuberculosis drug resistant strains from Russia. *PloS one*, 8(2), p.e56577. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3577857&tool=pmcentrez&rendertype=abstract> [Accessed March 26, 2014].
- Inouye, M. *et al.*, 2012. Short read sequence typing (SRST): multi-locus sequence types from short reads. *BMC genomics*, 13, p.338. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3460743&tool=pmcentrez&rendertype=abstract> [Accessed July 15, 2014].
- Ioerger, T.R. *et al.*, 2009. Genome analysis of multi- and extensively-drug-resistant tuberculosis from KwaZulu-Natal, South Africa. *PloS one*, 4(11), p.e7778. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2767505&tool=pmcentrez&rendertype=abstract> [Accessed August 15, 2011].
- Isaza, J.P. *et al.*, 2012. Whole genome shotgun sequencing of one Colombian clinical isolate of Mycobacterium tuberculosis reveals DosR regulon gene deletions. *FEMS microbiology letters*, 330(2), pp.113–20. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/22404577> [Accessed March 26, 2014].
- Jagielski, T. & Grzeszczuk, M., 2013. Identification and analysis of mutations in the katG gene in multidrug-resistant Mycobacterium tuberculosis clinical isolates. *Pneumonol. ...*, pp.298–307. Available at: <http://czasopisma.viamedica.pl/pap/article/download/34789/25395> [Accessed September 28, 2013].
- Jin, J. *et al.*, 2012. Evaluation of the GenoType® MTBDRplus assay and identification of a rare mutation for improving MDR-TB detection. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease*, 16(4), pp.521–6. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/22325117>.
- Jin, J. *et al.*, 2013. Underestimation of the resistance of Mycobacterium tuberculosis to second-line drugs by the new GenoType MTBDRsl test. *The Journal of molecular diagnostics : JMD*, 15(1), pp.44–50. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23159109> [Accessed September 30, 2013].
- Jnawali, H.N. *et al.*, 2013. Characterization of mutations in multi- and extensive drug resistance among strains of Mycobacterium tuberculosis clinical isolates in Republic of Korea. *Diagnostic microbiology and infectious disease*, 76(2), pp.187–96. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23561273> [Accessed September 28, 2013].

- Joshi, K.R., Dhiman, H. & Scaria, V., 2014. tbvar: a comprehensive genome variation resource for Mycobacterium tuberculosis. *Database : the journal of biological databases and curation*, 2014, p.bat083. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3885892&tool=pmcentrez&rendertype=abstract> [Accessed May 15, 2014].
- Joung, S., Jeon, S. & Lim, Y., 2013. Complete genome sequence of Mycobacterium bovis BCG Korea, the Korean vaccine strain for substantial production. *Genome Announcements*, 1(2), pp.87–88. Available at: <http://genomea.asm.org/content/1/2/e00069-13.short> [Accessed March 26, 2014].
- Kamerbeek, J. *et al.*, 1997. Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. *J. Clin. Microbiol.*, 35(4), pp.907–14. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=229700&tool=pmcentrez&rendertype=abstract> [Accessed October 25, 2011].
- Kato-Maeda, M. *et al.*, 2012. Beijing sublineages of Mycobacterium tuberculosis differ in pathogenicity in the guinea pig. *Clinical and vaccine immunology : CVI*, 19(8), pp.1227–37. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3416080&tool=pmcentrez&rendertype=abstract> [Accessed November 24, 2013].
- Kato-Maeda, M. & Gagneux, S., 2011. Strain classification of Mycobacterium tuberculosis: congruence between large sequence polymorphisms and spoligotypes. *INT J TUBERC LUNG DIS*, 15(July 2010), pp.131–133. Available at: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3600895/> [Accessed November 24, 2013].
- Kato-Maeda, M. & Kim, E., 2010. Differences among sublineages of the East-Asian lineage of Mycobacterium tuberculosis in genotypic clustering. ... *against Tuberculosis ...*, 14(June 2009), pp.538–544. Available at: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3625672/> [Accessed December 16, 2013].
- Kaufmann, S.H.E. *et al.*, 2014. Progress in tuberculosis vaccine development and host-directed therapies--a state of the art review. *The lancet. Respiratory medicine*, 2(4), pp.301–20. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24717627> [Accessed July 27, 2014].
- Kim, S. & Misra, A., 2007. SNP genotyping: technologies and biomedical applications. *Annual review of biomedical engineering*, 9, pp.289–320. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/17391067> [Accessed July 11, 2014].
- Köser, C.U. *et al.*, 2013. Rapid single-colony whole-genome sequencing of bacterial pathogens. *The Journal of antimicrobial chemotherapy*, pp.1–7. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24370932> [Accessed January 22, 2014].
- Köser, C.U. *et al.*, 2012. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS pathogens*, 8(8), p.e1002824. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3410874&tool=pmcentrez&rendertype=abstract> [Accessed January 12, 2014].
- Krzywinski, M. *et al.*, 2009. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9), pp.1639–1645. Available at: <http://genome.cshlp.org/content/19/9/1639.short> [Accessed October 28, 2011].
- Kurtz, S. *et al.*, 2004. Versatile and open software for comparing large genomes. *Genome Biol.*, 5(2), p.R12. Available at:

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=395750&tool=pmcentrez&rendertype=abstract>.

- Laabei, M., Recker, M. & Rudkin, J., 2014. Predicting the virulence of MRSA from its genome sequence. *Genome ...*, pp.839–849. Available at: <http://genome.cshlp.org/content/24/5/839.short> [Accessed June 30, 2014].
- Lacoma, a *et al.*, 2008. GenoType MTBDRplus assay for molecular detection of rifampin and isoniazid resistance in Mycobacterium tuberculosis strains and clinical samples. *Journal of clinical microbiology*, 46(11), pp.3660–7. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2576567&tool=pmcentrez&rendertype=abstract> [Accessed October 30, 2013].
- Lange, C. *et al.*, 2014. Management of patients with multidrug-resistant/extensively drug-resistant tuberculosis in Europe: a TBNET consensus statement. *The European respiratory journal*. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24659544> [Accessed May 2, 2014].
- Langmead, B. *et al.*, 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10(3), p.R25.3.
- Laurenzo, D. & Mousa, S. a, 2011. Mechanisms of drug resistance in Mycobacterium tuberculosis and current status of rapid molecular diagnostic testing. *Acta tropica*, 119(1), pp.5–10. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/21515239> [Accessed September 29, 2013].
- Layer, R.M. *et al.*, 2014. LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*, 15(6), p.R84. Available at: <http://genomebiology.com/2014/15/6/R84> [Accessed June 26, 2014].
- Lee, H. & Schatz, M.C., 2012. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics (Oxford, England)*, 28(16), pp.2097–105. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3413383&tool=pmcentrez&rendertype=abstract>.
- Lee, R.E. *et al.*, 2014. Spectinamides: a new class of semisynthetic antituberculosis agents that overcome native drug efflux. *Nature medicine*, 20(2), pp.152–8. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24464186> [Accessed June 4, 2014].
- Lew, J.M. *et al.*, 2012. Database resources for the tuberculosis community. *Tuberculosis (Edinb)*, pp.1–6. Available at: <http://dx.doi.org/10.1016/j.tube.2012.11.003>.
- Lew, J.M. *et al.*, 2011. TubercuList - 10 years after. *Tuberculosis*, 91, pp.1–7. Available at: <http://www.sciencedirect.com/science/article/pii/S1472979210001113>.
- Li, H., 2012. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics (Oxford, England)*, 28(14), pp.1838–44. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3389770&tool=pmcentrez&rendertype=abstract>.
- Li, H. *et al.*, 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), pp.2078–2079. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/19505943>.

- Li, H. & Durbin, R., 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, 26(5), pp.589–595. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/20080505>.
- Li, H. & Homer, N., 2010. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics*, 11(5), pp.473–83. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2943993&tool=pmcentrez&rendertype=abstract> [Accessed March 2, 2012].
- Li, H., Ruan, J. & Durbin, R., 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11), pp.1851–8. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2577856&tool=pmcentrez&rendertype=abstract>.
- Lin, P.L. & Flynn, J.L., 2010. Understanding latent tuberculosis: a moving target. *Journal of immunology (Baltimore, Md. : 1950)*, 185(1), pp.15–22. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3311959&tool=pmcentrez&rendertype=abstract> [Accessed December 15, 2013].
- Lin, S.-Y.G. *et al.*, 2013. Pyrosequencing for Rapid Detection of Extensively Drug-Resistant Tuberculosis in Clinical Isolates and Clinical Specimens. *Journal of clinical microbiology*, (November). Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24285731> [Accessed December 2, 2013].
- Ling, D.I., Zwerling, A.A. & Pai, M., 2008. GenoType MTBDR assays for the diagnosis of multidrug-resistant tuberculosis: a meta-analysis. *The European respiratory journal*, 32(5), pp.1165–74. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/18614561> [Accessed August 4, 2014].
- Liu, F. *et al.*, 2014. Comparative genomic analysis of Mycobacterium tuberculosis clinical isolates. *BMC genomics*, 15(1), p.469. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4070564&tool=pmcentrez&rendertype=abstract> [Accessed July 7, 2014].
- Liu, L. *et al.*, 2012. Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology*, 2012, pp.1–11. Available at: <http://www.hindawi.com/journals/jbb/2012/251364/> [Accessed July 6, 2012].
- Liu, Q. *et al.*, 2013. Triplex real-time PCR melting curve analysis for detecting Mycobacterium tuberculosis mutations associated with resistance to second-line drugs in a single reaction. *The Journal of antimicrobial chemotherapy*, 68(5), pp.1097–103. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23288402> [Accessed September 28, 2013].
- López, B. *et al.*, 2003. A marked difference in pathogenesis and immune response induced by different Mycobacterium tuberculosis genotypes. *Clinical and experimental immunology*, 133(1), pp.30–7. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1808750&tool=pmcentrez&rendertype=abstract>.
- Machado, D. *et al.*, 2012. Contribution of efflux to the emergence of isoniazid and multidrug resistance in Mycobacterium tuberculosis. *PloS one*, 7(4), p.e34538. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3321020&tool=pmcentrez&rendertype=abstract> [Accessed June 18, 2014].
- Machado, D. *et al.*, 2013. High-level resistance to isoniazid and ethionamide in multidrug-resistant Mycobacterium tuberculosis of the Lisboa family is associated with inhA double

- mutations. *The Journal of antimicrobial chemotherapy*, 68(8), pp.1728–32. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23539241> [Accessed September 28, 2013].
- Madhvilatha, G.K. *et al.*, 2012. Whole-genome sequences of two clinical isolates of *Mycobacterium tuberculosis* from Kerala, South India. *Journal of bacteriology*, 194(16), p.4430. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3416242&tool=pmcentrez&rendertype=abstract> [Accessed March 26, 2014].
- Magoc, T. *et al.*, 2013. GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics (Oxford, England)*, 29(14), pp.1718–25. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3702249&tool=pmcentrez&rendertype=abstract> [Accessed January 18, 2014].
- Malik, S. *et al.*, 2012. New insights into fluoroquinolone resistance in *Mycobacterium tuberculosis*: functional genetic analysis of *gyrA* and *gyrB* mutations. *PloS one*, 7(6), p.e39754. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3386181&tool=pmcentrez&rendertype=abstract> [Accessed April 3, 2014].
- Maruri, F. *et al.*, 2012. A systematic review of gyrase mutations associated with fluoroquinolone-resistant *Mycobacterium tuberculosis* and a proposed gyrase numbering system. *The Journal of antimicrobial chemotherapy*, 67(4), pp.819–31. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3299416&tool=pmcentrez&rendertype=abstract> [Accessed September 25, 2013].
- McEvoy, C.R.E. *et al.*, 2012. Comparative analysis of *Mycobacterium tuberculosis* *pe* and *ppe* genes reveals high sequence variation and an apparent absence of selective constraints. *PloS one*, 7(4), p.e30593. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3319526&tool=pmcentrez&rendertype=abstract> [Accessed November 1, 2012].
- McKenna, A. *et al.*, 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20(9), pp.1297–303. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2928508&tool=pmcentrez&rendertype=abstract> [Accessed May 22, 2013].
- Mdluli, K. *et al.*, 1998. Mechanisms involved in the intrinsic isoniazid resistance of *Mycobacterium avium*. *Molecular microbiology*, 27(6), pp.1223–33. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/9570407>.
- Migliori, G.B. *et al.*, 2009. MDR-TB and XDR-TB: drug resistance and treatment outcomes. *The European respiratory journal*, 34(3), pp.778–9. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/19720816> [Accessed May 17, 2014].
- Miotto, P. *et al.*, 2012. Genome-wide discovery of small RNAs in *Mycobacterium tuberculosis*. *PloS one*, 7(12), p.e51950. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3526491&tool=pmcentrez&rendertype=abstract> [Accessed July 10, 2014].
- Monahan, I.M. *et al.*, 2001. Differential expression of mycobacterial proteins following phagocytosis by macrophages. *Microbiology (Reading, England)*, 147(Pt 2), pp.459–71. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/11158363> [Accessed August 12, 2014].

- Morlock, G. & Metchock, B., 2003. ethA, inhA, and katG loci of ethionamide-resistant clinical Mycobacterium tuberculosis isolates. *Antimicrobial agents ...*, 47(12), pp.3799–3805. Available at: <http://aac.asm.org/content/47/12/3799.short> [Accessed October 4, 2013].
- Moure, R. *et al.*, 2013. Detection of streptomycin and quinolone resistance in Mycobacterium tuberculosis by a low-density DNA array. *Tuberculosis (Edinburgh, Scotland)*, 93(5), pp.508–14. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23906937> [Accessed September 28, 2013].
- Mukhopadhyay, S. & Balaji, K.N., 2011. The PE and PPE proteins of Mycobacterium tuberculosis. *Tuberculosis (Edinburgh, Scotland)*, 91(5), pp.441–447. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/21527209> [Accessed August 15, 2011].
- Nahid, P. *et al.*, 2010. Influence of M. tuberculosis lineage variability within a clinical trial for pulmonary tuberculosis. *PloS one*, 5(5), p.e10753. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2873999&tool=pmcentrez&rendertype=abstract> [Accessed November 9, 2013].
- Narayanan, S. & Deshpande, U., 2013. Whole-genome sequences of four clinical isolates of Mycobacterium tuberculosis from Tamil Nadu, south India. *Genome announcements*, 1(3), p.4430. Available at: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3707582/> [Accessed March 26, 2014].
- NCBI, 2014. Genome Assembly and Annotation report, Mycobacterium tuberculosis. Available at: <https://www.ncbi.nlm.nih.gov/genome/genomes/166>.
- Nebenzahl-Guimaraes, H. *et al.*, 2014. Systematic review of allelic exchange experiments aimed at identifying mutations that confer drug resistance in Mycobacterium tuberculosis. *The Journal of antimicrobial chemotherapy*, 69(2), pp.331–42. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24055765> [Accessed January 21, 2014].
- Ng, P.C. & Henikoff, S., 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13), pp.3812–3814. Available at: <http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkg509> [Accessed April 28, 2014].
- Nielsen, R. *et al.*, 2011. Genotype and SNP calling from next-generation sequencing data. *Nature reviews. Genetics*, 12(6), pp.443–51. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/21587300> [Accessed October 27, 2012].
- Niemann, S. *et al.*, 2009. Genomic diversity among drug sensitive and multidrug resistant isolates of Mycobacterium tuberculosis with identical DNA fingerprints. *PloS one*, 4(10), p.e7407. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2756628&tool=pmcentrez&rendertype=abstract> [Accessed July 1, 2011].
- Nübel, U. *et al.*, 2010. A timescale for evolution, population expansion, and spatial spread of an emerging clone of methicillin-resistant Staphylococcus aureus. *PLoS pathogens*, 6(4), p.e1000855. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2851736&tool=pmcentrez&rendertype=abstract> [Accessed May 27, 2014].
- Nunes-Alves, C. *et al.*, 2014. In search of a new paradigm for protective immunity to TB. *Nature reviews. Microbiology*, 12(4), pp.289–299. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24590243> [Accessed March 7, 2014].

- Orduña, P. *et al.*, 2011. Genomic and proteomic analyses of *Mycobacterium bovis* BCG Mexico 1931 reveal a diverse immunogenic repertoire against tuberculosis infection. *BMC genomics*, 12(1), p.493. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3199284&tool=pmcentrez&rendertype=abstract> [Accessed March 25, 2014].
- Ottenhoff, T.H.M. & Kaufmann, S.H.E., 2012. Vaccines against tuberculosis: where are we and where do we need to go? *PLoS pathogens*, 8(5), p.e1002607. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3349743&tool=pmcentrez&rendertype=abstract> [Accessed July 14, 2014].
- Ouellet, H., Johnston, J.B. & de Montellano, P.R.O., 2011. Cholesterol catabolism as a therapeutic target in *Mycobacterium tuberculosis*. *Trends in microbiology*, 19(11), pp.530–9. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3205253&tool=pmcentrez&rendertype=abstract> [Accessed May 27, 2014].
- Perdigão, J. *et al.*, 2013. Unraveling *Mycobacterium tuberculosis* genomic diversity and evolution in 2 Lisbon, Portugal, a highly drug resistant setting. *BMC genomics (in press)*.
- Pérez-Lago, L. *et al.*, 2014. Whole Genome Sequencing Analysis of Inpatient Microevolution in *Mycobacterium tuberculosis*: Potential Impact on the Inference of Tuberculosis Transmission. *The Journal of infectious diseases*, 209(1), pp.98–108. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23945373> [Accessed January 15, 2014].
- Peterson, J. *et al.*, 2002. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *Journal of Bacteriology*, 184(19), pp.5479–5490. Available at: <http://jlb.asm.org/content/184/19/5479.short> [Accessed March 26, 2014].
- Pooran, A. *et al.*, 2013. What is the cost of diagnosis and management of drug resistant tuberculosis in South Africa? *PloS one*, 8(1), p.e54587. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3548831&tool=pmcentrez&rendertype=abstract> [Accessed May 4, 2014].
- Preston, M.D. *et al.*, 2014. PlasmoView: a web-based resource to visualise global *Plasmodium falciparum* genomic variation. *The Journal of infectious diseases*, 209(11), pp.1808–15. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4017360&tool=pmcentrez&rendertype=abstract> [Accessed August 7, 2014].
- Randall, P.J. *et al.*, 2014. Neurons are host cells for *Mycobacterium tuberculosis*. *Infection and immunity*, 82(5), pp.1880–90. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24566619> [Accessed August 3, 2014].
- Rashdi, A. Al & Jadhav, B., 2014. Whole-genome sequencing and annotation of a clinical isolate of *Mycobacterium tuberculosis* from Mumbai, India. *Genome Announcements*, 2(2), p.2014. Available at: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3945509/> [Accessed May 11, 2014].
- Rausch, T. *et al.*, 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18), pp.i333–i339. Available at: <http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bts378> [Accessed September 7, 2012].
- Reddy, T.B.K. *et al.*, 2009. TB database: an integrated platform for tuberculosis research. *Nucleic acids research*, 37(Database issue), pp.D499–508. Available at:

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2686437&tool=pmcentrez&rendertype=abstract> [Accessed May 15, 2014].

- Reiling, N., Homolka, S. & Walter, K., 2013. Clade-Specific Virulence Patterns of *Mycobacterium tuberculosis*.
- Ren, H. & Liu, J., 2006. AsnB is involved in natural resistance of *Mycobacterium smegmatis* to multiple drugs. *Antimicrobial agents and chemotherapy*, 50(1), pp.250–255. Available at: <http://aac.asm.org/content/50/1/250.short> [Accessed June 28, 2014].
- Reyes, A. *et al.*, 2012. IS-seq: a novel high throughput survey of in vivo IS6110 transposition in multiple *Mycobacterium tuberculosis* genomes. *BMC genomics*, 13, p.249. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3443423&tool=pmcentrez&rendertype=abstract> [Accessed October 29, 2012].
- Rigouts, L. *et al.*, 2013. Rifampin resistance missed in automated liquid culture system for *Mycobacterium tuberculosis* isolates with specific *rpoB* mutations. *Journal of clinical microbiology*, 51(8), pp.2641–5. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3719602&tool=pmcentrez&rendertype=abstract> [Accessed May 15, 2014].
- Rodríguez, J. & Hernández, A., 2014. Global Adaptation to a Lipid Environment Triggers the Dormancy-Related Phenotype of *Mycobacterium tuberculosis*. *mBio*. Available at: <http://mbio.asm.org/content/5/3/e01125-14.short> [Accessed July 16, 2014].
- Rodwell, T.C. *et al.*, 2013. Predicting Extensively Drug-resistant Tuberculosis (XDR-TB) Phenotypes with Genetic Mutations. *Journal of clinical microbiology*, (December). Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24353002> [Accessed January 21, 2014].
- Roetzer, A. *et al.*, 2013. Whole Genome Sequencing versus Traditional Genotyping for Investigation of a *Mycobacterium tuberculosis* Outbreak: A Longitudinal Molecular Epidemiological Study. *PLoS Med.*, 10(2), p.e1001387.
- Safi, H. *et al.*, 2010. Allelic exchange and mutant selection demonstrate that common clinical *embCAB* gene mutations only modestly increase resistance to ethambutol in *Mycobacterium tuberculosis*. *Antimicrobial agents and chemotherapy*, 54(1), pp.103–8. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2798522&tool=pmcentrez&rendertype=abstract> [Accessed April 1, 2014].
- Safi, H. *et al.*, 2013. Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl- β -D-arabinose biosynthetic and utilization pathway genes. *Nature genetics*, 45(10), pp.1190–1197. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23995136> [Accessed September 17, 2013].
- Sandgren, A. *et al.*, 2009. Tuberculosis drug resistance mutation database. *PLoS Med.*, 6(2), p.e1000002. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2637921&tool=pmcentrez&rendertype=abstract> [Accessed November 5, 2012].
- Satou, K. *et al.*, 2014. Complete Genome Sequences of Eight *Helicobacter pylori* Strains with Different Virulence Factor Genotypes and Methylation Profiles, Isolated from Patients with Diverse Gastrointestinal Diseases on Okinawa Island, Japan, Determined Using PacBio Single-Molec. *Genome announcements*, 2(2). Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3990747&tool=pmcentrez&rendertype=abstract> [Accessed July 16, 2014].

- Sboner, A. *et al.*, 2011. The real cost of sequencing: higher than you think! *Genome Biol.*, 12(8), p.125. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245608&tool=pmcentrez&rendertype=abstract>.
- Schliep, K.P., 2011. phangorn: phylogenetic analysis in R. *Bioinformatics (Oxford, England)*, 27(4), pp.592–3. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3035803&tool=pmcentrez&rendertype=abstract> [Accessed May 26, 2014].
- Schürch, A.C. *et al.*, 2010. The tempo and mode of molecular evolution of Mycobacterium tuberculosis at patient-to-patient scale. *Infect. Genet. Evol.*, 10(1), pp.108–14. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/19835997> [Accessed August 16, 2013].
- Schürch, A.C. & van Soolingen, D., 2012. DNA fingerprinting of Mycobacterium tuberculosis: from phage typing to whole-genome sequencing. *Infect. Genet. Evol.*, 12(4), pp.602–9. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/22067515> [Accessed November 1, 2012].
- Scorpio, A. & Zhang, Y., 1996. Mutations in pncA, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus. *Nature medicine*, 2(6), pp.662–7. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/8640557> [Accessed May 13, 2014].
- Seki, M. *et al.*, 2009. Whole genome sequence analysis of Mycobacterium bovis bacillus Calmette-Guérin (BCG) Tokyo 172: a comparative study of BCG vaccine substrains. *Vaccine*, 27(11), pp.1710–6. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/19200449> [Accessed July 24, 2012].
- Sekiguchi, J.-I. *et al.*, 2007. Development and evaluation of a line probe assay for rapid identification of pncA mutations in pyrazinamide-resistant mycobacterium tuberculosis strains. *Journal of clinical microbiology*, 45(9), pp.2802–7. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2045301&tool=pmcentrez&rendertype=abstract> [Accessed September 29, 2013].
- Sengstake, S. *et al.*, 2014. Optimizing multiplex SNP-based data analysis for genotyping of Mycobacterium tuberculosis isolates. *BMC genomics*, 15(1), p.572. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/25001491>.
- Shabbeer, A. *et al.*, 2012. Web tools for molecular epidemiology of tuberculosis. *Infect. Genet. Evol.*, 12(4), pp.767–81. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/21903179> [Accessed October 26, 2012].
- Sharon J. Peacock, P.D., 2013. Whole-Genome Sequencing for Rapid Susceptibility Testing of M. tuberculosis. *The New England Journal of Medicine*.
- Shean, K. *et al.*, 2013. Drug-associated adverse events and their relationship with outcomes in patients receiving treatment for extensively drug-resistant tuberculosis in South Africa. *PloS one*, 8(5), p.e63057. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3646906&tool=pmcentrez&rendertype=abstract> [Accessed May 4, 2014].
- Shi, W. *et al.*, 2011. Pyrazinamide inhibits trans-translation in Mycobacterium tuberculosis. *Science (New York, N.Y.)*, 333(6049), pp.1630–2. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3502614&tool=pmcentrez&rendertype=abstract> [Accessed May 4, 2014].

- Shi, X. *et al.*, 2013. Development of a single multiplex amplification refractory mutation system PCR for the detection of rifampin-resistant *Mycobacterium tuberculosis*. *Gene*, 530(1), pp.95–9. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23968719> [Accessed September 28, 2013].
- Shitikov, E. a. *et al.*, 2014. Unusual Large-Scale Chromosomal Rearrangements in *Mycobacterium tuberculosis* Beijing B0/W148 Cluster Isolates P. Supply, ed. *PLoS ONE*, 9(1), p.e84971. Available at: <http://dx.plos.org/10.1371/journal.pone.0084971> [Accessed January 12, 2014].
- Slayden, R. a & Barry, C.E., 2000. The genetics and biochemistry of isoniazid resistance in *mycobacterium tuberculosis*. *Microbes and infection / Institut Pasteur*, 2(6), pp.659–69. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/10884617>.
- Smigielski, E.M. *et al.*, 2000. dbSNP: a database of single nucleotide polymorphisms. *Nucleic acids research*, 28(1), pp.352–5. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102496&tool=pmcentrez&rendertype=abstract>.
- Smits, S. a & Ouverney, C.C., 2010. jsPhyloSVG: a javascript library for visualizing interactive and vector-based phylogenetic trees on the web. *PloS one*, 5(8), p.e12267. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2923619&tool=pmcentrez&rendertype=abstract> [Accessed March 27, 2012].
- Stamatakis, A., Hoover, P. & Rougemont, J., 2008. A rapid bootstrap algorithm for the RAxML Web servers. *Syst. Biol.*, 57(5), pp.758–71. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/18853362> [Accessed February 27, 2013].
- Steingart, K.R. *et al.*, 2014. Xpert® MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults. *The Cochrane database of systematic reviews*, 1, p.CD009593. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24448973> [Accessed August 13, 2014].
- Stoffels, K. *et al.*, 2012a. Systematic analysis of pyrazinamide-resistant spontaneous mutants and clinical isolates of *Mycobacterium tuberculosis*. *Antimicrobial agents and chemotherapy*, 56(10), pp.5186–93. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3457413&tool=pmcentrez&rendertype=abstract> [Accessed September 30, 2013].
- Stoffels, K. *et al.*, 2012b. Systematic analysis of pyrazinamide-resistant spontaneous mutants and clinical isolates of *Mycobacterium tuberculosis*. *Antimicrobial agents and chemotherapy*, 56(10), pp.5186–93. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/22825123> [Accessed November 10, 2012].
- Stop TB Partnership, 2014. *Mycobacteriology Laboratory Manual*. , pp.51–66. Available at: http://www.stoptb.org/wg/gli/assets/documents/gli_mycobacteriology_lab_manual_web.pdf
- Stucki, D. *et al.*, 2012. Two new rapid SNP-typing methods for classifying *Mycobacterium tuberculosis* complex into the main phylogenetic lineages. *PloS one*, 7(7), p.e41253. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3401130&tool=pmcentrez&rendertype=abstract> [Accessed November 22, 2013].
- Stucki, D. & Gagneux, S., 2012. Single nucleotide polymorphisms in *Mycobacterium tuberculosis* and the need for a curated database. *Tuberculosis (Edinb)*, 30(1), pp.30–9.

Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23266261> [Accessed December 26, 2012].

- Supply, P. *et al.*, 2001. Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. *J. Clin. Microbiol.*, 39(10), pp.3563–71. Available at: <http://jcm.asm.org/cgi/content/abstract/39/10/3563> [Accessed October 11, 2011].
- Supply, P. *et al.*, 2013. Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*. *Nature genetics*, 45(2), pp.172–9. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3856870&tool=pmcentrez&rendertype=abstract> [Accessed January 10, 2014].
- Tahaoğlu, K. *et al.*, 2001. The treatment of multidrug-resistant tuberculosis in Turkey. *The New England journal of medicine*, 345(3), pp.170–4. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/11463011> [Accessed May 17, 2014].
- Takiff, H.E. *et al.*, 1994. Cloning and nucleotide sequence of *Mycobacterium tuberculosis* gyrA and gyrB genes and detection of quinolone resistance mutations. *Antimicrobial agents and chemotherapy*, 38(4), pp.773–80. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=284541&tool=pmcentrez&rendertype=abstract> [Accessed May 18, 2014].
- Tan, Y. *et al.*, 2013. Role of pncA and rpsA Gene Sequencing in Diagnosis of Pyrazinamide Resistance in *Mycobacterium tuberculosis* Isolates from Southern China. *Journal of clinical microbiology*, (October). Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24131688> [Accessed October 22, 2013].
- Telenti, A. *et al.*, 1997. The emb operon, a gene cluster of *Mycobacterium tuberculosis* involved in resistance to ethambutol. *Nature medicine*, 3(5), pp.567–70. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/9142129> [Accessed May 17, 2014].
- Tessema, B. *et al.*, 2013. Molecular epidemiology and transmission dynamics of *Mycobacterium tuberculosis* in Northwest Ethiopia: new phylogenetic lineages found in Northwest Ethiopia. *BMC infectious diseases*, 13(1), p.131. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3605317&tool=pmcentrez&rendertype=abstract> [Accessed February 26, 2014].
- Thwaites, G. *et al.*, 2008. Relationship between *Mycobacterium tuberculosis* genotype and the clinical phenotype of pulmonary and meningeal tuberculosis. *Journal of clinical microbiology*, 46(4), pp.1363–8. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2292951&tool=pmcentrez&rendertype=abstract> [Accessed November 15, 2013].
- Timp, W. & Mirsaidov, U., 2010. Nanopore sequencing: electrical measurements of the code of life. *Nanotechnology*, ..., 9(3), pp.281–294. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5422648 [Accessed January 24, 2014].
- Uplekar, S. *et al.*, 2013. High-resolution transcriptome and genome-wide dynamics of RNA polymerase and NusA in *Mycobacterium tuberculosis*. *Nucleic acids research*, 41(2), pp.961–77. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3553938&tool=pmcentrez&rendertype=abstract> [Accessed July 16, 2014].

- Utturkar, S.M. *et al.*, 2014. Evaluation and validation of de novo and hybrid assembly techniques to derive high quality genome sequences.
- Vincent, V. *et al.*, 2012. The TDR Tuberculosis Strain Bank: a resource for basic science, tool development and diagnostic services. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease*, 16(1), pp.24–31. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/22236841>.
- Vishnoi, A. *et al.*, 2008. MGDD: Mycobacterium tuberculosis genome divergence database. *BMC genomics*, 9, p.373. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2518163&tool=pmcentrez&rendertype=abstract> [Accessed May 15, 2014].
- De Vos, M. *et al.*, 2013. Putative compensatory mutations in the rpoC gene of rifampin-resistant Mycobacterium tuberculosis are associated with ongoing transmission. *Antimicrobial agents and chemotherapy*, 57(2), pp.827–32. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3553702&tool=pmcentrez&rendertype=abstract> [Accessed May 17, 2014].
- Waddell, S.J. & Butcher, P.D., 2010. Use of DNA arrays to study transcriptional responses to antimycobacterial compounds. *Methods in molecular biology (Clifton, N.J.)*, 642, pp.75–91. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/20401587> [Accessed August 12, 2014].
- Walker, T.M. *et al.*, 2014. Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007–12, with whole pathogen genome sequences: an observational study. *The Lancet Respiratory Medicine*, pp.285–292. Available at: <http://linkinghub.elsevier.com/retrieve/pii/S221326001470027X> [Accessed March 20, 2014].
- Walker, T.M. *et al.*, 2013. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet infectious diseases*, 13(2), pp.137–46. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3556524&tool=pmcentrez&rendertype=abstract> [Accessed February 23, 2014].
- Wang, J. *et al.*, 2011. CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods*, 8(8), pp.652–656. Available at: <http://www.nature.com/nmeth/journal/v8/n8/abs/nmeth.1628.html> [Accessed April 19, 2012].
- Wang, X. *et al.*, 2013. A simple, rapid and economic method for detecting multidrug-resistant tuberculosis. *The Brazilian journal of infectious diseases : an official publication of the Brazilian Society of Infectious Diseases*, (x x), pp.2–6. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24029439> [Accessed September 24, 2013].
- Warner, D.F. & Mizrahi, V., 2013. Complex genetics of drug resistance in Mycobacterium tuberculosis. *Nature Genetics*, 45(10), pp.1107–1108. Available at: <http://www.nature.com/doifinder/10.1038/ng.2769> [Accessed September 26, 2013].
- Wattam, A.R. *et al.*, 2014. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic acids research*, 42(Database issue), pp.D581–91. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965095&tool=pmcentrez&rendertype=abstract> [Accessed April 29, 2014].
- Wolf, A.J. *et al.*, 2008. Initiation of the adaptive immune response to Mycobacterium tuberculosis depends on antigen production in the local lymph node, not the lungs. *The*

- Journal of experimental medicine*, 205(1), pp.105–15. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2234384&tool=pmcentrez&rendertype=abstract> [Accessed July 14, 2014].
- Wolf, A.J. *et al.*, 2007. Mycobacterium tuberculosis infects dendritic cells with high frequency and impairs their function in vivo. *Journal of immunology (Baltimore, Md. : 1950)*, 179(4), pp.2509–19. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/17675513> [Accessed August 3, 2014].
- World Health Organization, 2013. *Global tuberculosis report 2013*,
- Ye, K. *et al.*, 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21), pp.2865–71. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2781750&tool=pmcentrez&rendertype=abstract> [Accessed March 9, 2012].
- Yee, D. *et al.*, 2003. Incidence of serious side effects from first-line antituberculosis drugs among patients treated for active tuberculosis. *American journal of respiratory and critical care medicine*, 167(11), pp.1472–7. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/12569078> [Accessed May 17, 2014].
- Yuen, K.Y. *et al.*, 1995. IS6110 based amplotyping assay and RFLP fingerprinting of clinical isolates of Mycobacterium tuberculosis. *J. Clin. Pathol.*, 48(10), pp.924–8. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=502948&tool=pmcentrez&rendertype=abstract>.
- Zaharia, M. *et al.*, 2011. Faster and More Accurate Sequence Alignment with SNAP. *Arxiv.org*. Available at: <http://arxiv.org/abs/1111.5572> [Accessed August 4, 2014].
- Zerbino, D.R. & Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, 18(5), pp.821–9. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2336801&tool=pmcentrez&rendertype=abstract> [Accessed July 16, 2011].
- Zhang, H. *et al.*, 2013. Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance. *Nature genetics*, 45(10), pp.1255–1260. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23995137> [Accessed September 17, 2013].
- Zhang, S. *et al.*, 2013. Mutations in panD encoding aspartate decarboxylase are associated with pyrazinamide resistance in Mycobacterium tuberculosis. *Emerging Microbes & Infections*, 2(6), p.e34. Available at: <http://www.nature.com/doifinder/10.1038/emi.2013.38> [Accessed November 5, 2013].
- Zhang, W. *et al.*, 2011. A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PLoS one*, 6(3), p.e17915. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3056720&tool=pmcentrez&rendertype=abstract> [Accessed July 18, 2011].
- Zhang, Y. *et al.*, 2011. Complete genome sequences of Mycobacterium tuberculosis strains CCDC5079 and CCDC5080, which belong to the Beijing family. *Journal of bacteriology*, 193(19), pp.5591–2. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3187387&tool=pmcentrez&rendertype=abstract> [Accessed March 26, 2014].

- Zhang, Y. & Vilcheze, C., 2009. Mechanisms of drug resistance in *Mycobacterium tuberculosis*. *INT J TUBERC LUNG DIS*, 13(11), pp.1320–1330. Available at: <http://books.google.com/books?hl=en&lr=&id=il9iFALTd8cC&oi=fnd&pg=PA115&dq=Mechanisms+of+drug+resistance+in+Mycobacterium+tuberculosis∓ots=77KhfYn0LA&sig=Zhw8QtjaE18HwmTqAT1Mn54Ao6s> [Accessed July 6, 2012].
- Zhao, F. *et al.*, 2014. Binding pocket alterations in dihydrofolate synthase confer resistance to para-aminosalicylic acid in clinical isolates of *Mycobacterium tuberculosis*. *Antimicrobial agents and chemotherapy*, 58(3), pp.1479–87. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24366731> [Accessed July 6, 2014].
- Zimenkov, D. V *et al.*, 2013. Detection of second-line drug resistance in *Mycobacterium tuberculosis* using oligonucleotide microarrays. *BMC infectious diseases*, 13(1), p.240. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3671172&tool=pmcentrez&rendertype=abstract> [Accessed September 28, 2013].
- Zumla, A.I. *et al.*, 2014. New antituberculosis drugs, regimens, and adjunct therapies: needs, advances, and future prospects. *The Lancet infectious diseases*, 14(4), pp.327–40. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24670627> [Accessed July 10, 2014].

SUPPLEMENTARY MATERIAL

Supplementary Table 1 *Mtb* complete genomes used to generate a set of validated SNP, indel and large deletion loci.

Genome name (Genbank Accession Number)	SNPs (n)	SNPs (% observed)	Indels (n)	Indels (% observed)	Large Del. (n)	Large Del. (% observed)
7199, 99 (NC_020089)	1,013	72.75	108	70.37	12	41.67
CAS/NITR204 (NC_021193)	4,620	23.79	1,963	4.12	2	100
CCDC5079 (NC_017523)	2,131	60.39	385	25.97	11	54.54
CCDC5180 (NC_017522)	1,751	71.50	185	55.67	13	46.15
CDC1551 (NC_002755)	1,222	66.77	207	36.71	15	46.67
CTRI, 2 (CP002992)	983	79.14	99	69.70	14	42.86
EAI5/NITR206 (NC_021194)	2,187	58.71	251	29.08	11	72.73
F11 (NC_009565)	986	80.16	92	75.00	8	37.50
KZN 1435 (NC_012943)	1,005	78.50	113	55.75	7	42.86
KZN 4207 (NC_016768)	994	79.48	93	67.74	6	66.67
KZN 605 (NC_018078)	1,019	78.02	112	56.25	6	50.00
RGTB327 (CP003233)	1,145	56.33	1,821	2.58	7	0.00
RGTB423 (NC_017528)	2,615	71.59	2,027	4.53	4	0.00
Beijing/NITR203 (NC_021054)	2,338	56.97	198	39.89	0	0.00
Erdman (AP012340)	1,159	70.06	114	61.40	25	32.00
UT205 (NC016934)	808	79.83	94	62.76	13	69.23
Overall	12,887		6,749		95	

Summary of genetic variation extracted for a set of 16 *Mtb* complete genomes downloaded from Genbank. Genetic variation across all 16 genomes was derived with respect to the H37Rv reference genome (Genbank accession number NC_000962.3). The number of SNPs, indels and large deletions per genome are shown in columns 2, 4 and 6 respectively. The percentage of variants also observed in the WGS public dataset (namely present in at least one of the 1,470 isolates in the WGS data set 1 after sample filtering) are indicated in columns 3, 5 and 7. Overall values represent the total number of variant sites discovered across all 16 genomes.

Supplementary Table 2 Experimental spoligotyping and *SpolPred* results for the 51 Ugandan samples

Isolate number	Experimental Results			SpolPred Results		
	Octal Code	SIT number	Spoligotype	Octal Code	SIT number	Spoligotype
18	47777777413771	126	EAI5	47777777413771	126	EAI5
48	77774677760771	1721	X1	77775677760771	302	X1
47	77775677760771	302	X1	77775677760771	302	X1
10	71777677760771	2356	X1	71777677760771	2356	X1
30	71777677760771	2356	X1	71777677760771	2356	X1
38	77777477760771	451	T-H37Rv	77777477760771	451	T-H37Rv
25	777767601560751	-	O	77777603560771	228	T1
35	77777607760771	42	LAM9	77777607760771	42	LAM9
27	77777606060771	59	LAM11_ZWE	77777606060771	59	LAM11_ZWE
34	77777606060771	59	LAM11_ZWE	77777606060771	59	LAM11_ZWE
42	77777606060771	59	LAM11_ZWE	77777606060771	59	LAM11_ZWE
36	77777606060771	59	LAM11_ZWE	77777606060771	59	LAM11_ZWE
2	Not determined	-	-	777737606060760	-	O
1	Not determined	-	-	777737606060760	-	O
3	Not determined	-	-	777737606060760	-	O
4	Not determined	-	-	777737606060760	-	O
17	77577606060731	1549	LAM11_ZWE	77577606060731	1549	LAM11_ZWE
7	Not determined	-	-	77577606060731	1549	LAM11_ZWE
31	77777777760771	53	T1	77777777760771	53	T1
16	77777777760771	53	T1	77777777760771	53	T1
39	77777777760771	53	T1	77777777760771	53	T1
43	77777777760771	53	T1	77777777760771	53	T1
20	000000007760771	-	O	000000007760771	-	O
51	000000007760731	125	LAM3	000000007760731	125	LAM3
5	000000007760771	-	O	000000007760771	-	O
6	000000007760771	-	O	000000007760771	-	O
37	63777477760730	420	T2-Uganda	63777477760730	420	T2-Uganda
45	77777777760731	52	T2	77777777760731	52	T2
9	Not determined	-	-	77777775760731	2867	T2
11	77777777760731	52	T2	77777777760731	52	T2
22	77777777760731	52	T2	77777777760731	52	T2
23	77777777760731	52	T2	77777777760731	52	T2
44	77777777760731	52	T2	77777777760731	52	T2
24	77777777760731	52	T2	77777777760731	52	T2
46	77777403760731	-	O	77777403760731	-	O
28	77777403760731	-	O	77777403760731	-	O
41	00000000003771	1	BEIJING	00000000003771	1	BEIJING

40	703767600001771	-	O	703777600001771	356	CAS1_DELHI
32	703777740003771	26	CAS1_DELHI	703777740003771	26	CAS1_DELHI
29	703777740003771	26	CAS1_DELHI	703777740003771	26	CAS1_DELHI
33	703777740003771	26	CAS1_DELHI	703777740003771	26	CAS1_DELHI
14	703777740003771	26	CAS1_DELHI	703777740003771	26	CAS1_DELHI
12	703777740003771	26	CAS1_DELHI	703777740003771	26	CAS1_DELHI
13	703777740003771	26	CAS1_DELHI	703777740003771	26	CAS1_DELHI
21	703377400001771	21	CAS1_KILI	703377400001771	21	CAS1_KILI
26	703367400001771	1675	CAS1_KILI	703377400001771	21	CAS1_KILI
8	700377740003771	288	CAS2	700377740003771	288	CAS2
49	700367700003771	-	O	700377740003771	288	CAS2
50	700377740003771	288	CAS2	700377740003771	288	CAS2
15	700377740003771	288	CAS2	700377740003771	288	CAS2
16	Not determined	-	-	700377740003771	288	CAS2

Experimental spoligotype results and *SpolPred* predicted ones for 51 Ugandan MTBC isolates.

Supplementary Table 3 Lineage composition of the WGS data set 2 populations

Study	<i>n</i>	# SNPs	Lin. 1 %	Lin. 2 %	Lin. 3 %	Lin. 4 %	Lin. 5 %	Lin. 6 %	Lin. 7 %	<i>M. bovis</i> %
Canada (Gardy <i>et al.</i> 2011)	19	1,021	0	0	0	100	0	0	0	0
China (H. Zhang <i>et al.</i> 2013)	161	19,314	0	75.8	1.2	23.0	0	0	0	0
Global (Comas <i>et al.</i> 2013)	166	30,770	14.5	34.9	12.7	24.1	7.8	5.4	0	0.6
Malawi (Guerra-Assunção <i>et al.</i> 2014)	338	19,240	18.9	5.6	14.8	60.7	0	0	0	0
Netherlands (Bryant, Schürch, <i>et al.</i> 2013)	125	8,635	4.0	16.8	0.8	78.4	0	0	0	0
Portugal (Perdigão <i>et al.</i> 2013)	81	7,163	0	6.2	1.2	92.6	0	0	0	0
Russia (Casali & Nikolayevskyy 2012)	259	18,699	1.5	58.7	1.9	35.9	0.4	0	0	1.5
Uganda (Clark <i>et al.</i> 2013)	51	8,019	2.0	2.0	27.5	68.6	0	0	0	0
UK (Walker <i>et al.</i> 2013)	390	19,408	5.9	2.6	24.4	65.1	0.5	0	0	1.5
Ethiopia (Firdessa <i>et al.</i> 2013)	4	2345	0	0	0	0	0	0	100	0
Djibouti (Blouin)	7	5445	0	28.6	0	0	14.3	28.6	28.6	0

<i>et al.</i> 2012)										
Overall	1,601	91648	7.5	24.4	11.8	53.4	1.1	0.7	0.4	0.7

Supplementary Table 4 The set of 62 phylogenetically informative SNPs for MTBC typing

lineage	Position*	Gene coor.	NT chg.	Codon num.	Codon chg.	AA chg.	Locus Id	Gene name
lineage1	615938	1104	G/A	368	GAG/GAA	E/E	Rv0524	<i>hemL</i>
lineage1.1	4404247	1056	G/A	352	CTG/CTA	L/L	Rv3915	-
lineage1.1.1	3021283	711	G/A	237	CGG/CGA	R/R	Rv2707	-
lineage1.1.1.1	3216553	339	G/A	113	GTC/GTT	V/V	Rv2907c	<i>rimM</i>
lineage1.1.2	2622402	51	G/A	17	GCC/GCT	A/A	Rv2343c	<i>dnaG</i>
lineage1.1.3	1491275	1038	G/A	346	CAC/CAT	H/H	Rv1326c	<i>glgB</i>
lineage1.2.1	3479545	375	C/A	125	GCC/GCA	A/A	Rv3111	<i>moaC1</i>
lineage1.2.2	3470377	303	C/T	101	CAG/CAA	Q/Q	Rv3101c	<i>ftsX</i>
lineage2	497491	810	G/A	270	GAC/GAT	D/D	Rv0411c	<i>glnH</i>
lineage2.1	1881090	5787	C/T	1929	GGC/GGT	G/G	Rv1661	<i>pks7</i>
lineage2.2	2505085	615	G/A	205	GCC/GCT	A/A	Rv2231c	<i>cobC</i>
lineage2.2.1	797736	804	C/T	268	CTC/CTT	L/L	Rv0697	-
lineage2.2.1.1	4248115	1602	C/T	534	GAC/GAT	D/D	Rv3795	<i>embB</i>
lineage2.2.1.2	3836274	618	G/A	206	TTC/TTT	F/F	Rv3417c	<i>groEL1</i>
lineage2.2.2	346693	1059	G/T	353	TCG/TCT	S/S	Rv0284	<i>eccC3</i>
lineage3	3273107	894	C/A	298	GCC/GCA	A/A	Rv2936	<i>drrA</i>
lineage3.1.1	1084911	840	G/A	280	TAC/TAT	Y/Y	Rv0973c	<i>accA2</i>
lineage3.1.2	3722702	930	G/C	310	CTC/CTG	L/L	Rv3336c	<i>trpS</i>
lineage3.1.2.1	1237818	375	C/G	125	CTG/CTC	L/L	Rv1111c	-
lineage3.1.2.2	2874344	2142	G/A	714	CGC/CGT	R/R	Rv2555c	<i>alaS</i>
lineage4**	931123	171	T/C	57	TAT/TAC	Y/Y	Rv0835	<i>lpqQ</i>
lineage4.1	62657	2262	G/A	754	CCG/CCA	P/P	Rv0058	<i>dnaB</i>
lineage4.1.1	514245	1077	C/T	359	GTG/GTA	V/V	Rv0425c	<i>ctpH</i>
lineage4.1.1.1	1850119	1917	C/T	639	ACG/ACA	T/T	Rv1640c	<i>lysX</i>
lineage4.1.1.2	541048	444	T/G	148	TCA/TCC	S/S	Rv0450c	<i>mmpL4</i>
lineage4.1.1.3	4229087	741	C/T	247	AAC/AAT	N/N	Rv3782	<i>glfT1</i>
lineage4.1.2	891756	514	A/G	172	TTG/CTG	L/L	Rv0798c	<i>cfp29</i>
lineage4.1.2.1	107794	195	C/T	65	GCC/GCT	A/A	Rv0098	<i>fcoT</i>
lineage4.2	2411730	393	G/C	131	TCC/TCG	S/S	Rv2152c	<i>murC</i>
lineage4.2.1	783601	1117	A/C	373	AGG/CGG	R/R	Rv0684	<i>fusA1</i>
lineage4.2.2	1487796	636	C/A	212	ATC/ATA	I/I	Rv1324	-
lineage4.2.2.1	1455780	286	T/C	96	TTG/CTG	L/L	Rv1299	<i>prfA</i>
lineage4.3	764995	1626	C/G	542	GCC/GCG	A/A	Rv0668	<i>rpoC</i>
lineage4.3.1	615614	780	C/A	260	GCC/GCA	A/A	Rv0524	<i>hemL</i>

lineage4.3.2	4316114	483	G/A	161	GCC/GCT	A/A	Rv3843c	-
lineage4.3.2.1	3388166	705	C/G	235	ACG/ACC	T/T	Rv3029c	<i>fixA</i>
lineage4.3.3	403364	2478	G/A	826	CCC/CCT	P/P	Rv0338c	-
lineage4.3.4	3977226	165	G/A	55	TTG/TTA	L/L	Rv3538	-
lineage4.3.4.1	4398141	1545	G/A	515	TCG/TCA	S/S	Rv3910	-
lineage4.3.4.2	1132368	744	C/T	248	ACC/ACT	T/T	Rv1013	<i>pks16</i>
lineage4.3.4.2.1	1502120	522	C/A	174	ACC/ACA	T/T	Rv1333	-
lineage4.4	4307886	1029	G/A	343	CGC/CGT	R/R	Rv3834c	<i>serS</i>
lineage4.4.1	4151558	660	G/A	220	GGC/GGT	G/G	Rv3708c	<i>asd</i>
lineage4.4.1.1	355181	684	G/A	228	AAG/AAA	K/K	Rv0291	<i>mycP3</i>
lineage4.4.1.2	2694560	405	G/C	135	CTC/CTG	L/L	Rv2397c	<i>cysA1</i>
lineage4.4.2	4246508	3276	G/A	1092	GCG/GCA	A/A	Rv3794	<i>embA</i>
lineage4.5	1719757	1032	G/T	344	CCG/CCT	P/P	Rv1524	-
lineage4.6	3466426	666	G/A	222	GTC/GTT	V/V	Rv3097c	<i>lipY</i>
lineage4.6.1	4260268	879	G/C	293	GCC/GCG	A/A	Rv3800c	<i>pks13</i>
lineage4.6.1.1	874787	555	G/A	185	CCG/CCA	P/P	Rv0781	<i>ptrBa</i>
lineage4.6.1.2	1501468	543	G/C	181	CCG/CCC	P/P	Rv1332	-
lineage4.6.2	4125058	642	G/C	214	CGG/CGC	R/R	Rv3683	-
lineage4.6.2.1	3570528	684	C/G	228	CGG/CGC	R/R	Rv3198c	<i>uvrD2</i>
lineage4.6.2.2	2875883	603	C/T	201	CTG/CTA	L/L	Rv2555c	<i>alaS</i>
lineage4.7	4249732	3219	C/G	1073	GCC/GCG	A/A	Rv3795	<i>embB</i>
lineage4.8	3836739	153	G/A	51	GAC/GAT	D/D	Rv3417c	<i>groEL1</i>
lineage4.9**	1759252	1572	G/T	524	TCG/TCT	S/S	Rv1552	<i>frdA</i>
lineage5	1799921	339	C/A	113	GGC/GGA	G/G	Rv1599	<i>hisD</i>
lineage6	1816587	399	C/G	133	GTC/GTG	V/V	Rv1617	<i>pykA</i>
lineage7	1137518	543	G/A	181	AAC/AAT	N/N	Rv1018c	<i>glmU</i>
lineageBOV	2831482	1110	A/G	370	GGT/GGC	G/G	Rv2515c	-
lineageBOV_AFRI	1882180	477	C/T	159	GCC/GCT	A/A	Rv1662	<i>pks8</i>

Minimum set of 62 SNPs for MTBC typing. SNPs are annotated using their chromosome coordinate on the H37Rv reference NC_000962.3 (*). The alternative allele (right hand nucleotide in column four) is the one specific to the lineage or sub-lineage, with the exception of two cases (**), lineage 4 and 4.9, in which the reference allele is the one specific to the lineage. SNPs are annotated providing the locus tag, gene name, gene coordinate, codon number, nucleotide change, codon change and resulting amino acid change.

Supplementary Table 5 Lineage predictions for a set of reference genomes

Species	Strain	Accession	Predicted lineage and sub-lineage*	Known lineage, sub-lineage (reference)**
<i>M. africanum</i>	GM041182	NC_015758.1	6, BOV_AFRI	Mycobacterium africanum West African 2 (Bentley <i>et al.</i> 2012)
<i>M. bovis</i>	BCG strain Korea 1168P	CP003900.2	BOV, BOV_AFRI	Mycobacterium bovis BCG (Joung <i>et al.</i> 2013)
<i>M. bovis</i>	BCG strain Mexico	NC_016804.1	BOV, BOV_AFRI	Mycobacterium bovis BCG (Orduña <i>et al.</i> 2011)
<i>M. bovis</i>	BCG strain Moreau RDJ	AM412059.2	BOV, BOV_AFRI	Mycobacterium bovis BCG (Gomes <i>et al.</i> 2011)
<i>M. bovis</i>	BCG strain Pasteur 1173P2	NC_008769.1	BOV, BOV_AFRI	Mycobacterium bovis BCG (Brosch <i>et al.</i> 2007)
<i>M. bovis</i>	BCG strain Tokyo 172	NC_012207.1	BOV, BOV_AFRI	Mycobacterium bovis BCG (Seki <i>et al.</i> 2009)
<i>M. tuberculosis</i>	7199, 99	NC_020089.1	4, 4.1, 4.1.2, 4.1.2.1	Hamburg clone, Haarlem lineage (Roetzer <i>et al.</i> 2013)
<i>M. tuberculosis</i>	BT1	CP002883.1	2, 2.2, 2.2.1	Beijing/W
<i>M. tuberculosis</i>	BT2	CP002882.1	2, 2.2, 2.2.1	Beijing/W
<i>M. tuberculosis</i>	CAS NITR204	NC_021193.1	3	CAS strain (Narayanan & Deshpande 2013)
<i>M. tuberculosis</i>	CCDC5079	NC_017523.1	2, 2.2, 2.2.1	Beijing family (Y. Zhang <i>et al.</i> 2011)
<i>M. tuberculosis</i>	CCDC5180	NC_017522.1	2, 2.2, 2.2.1	Beijing family (Y. Zhang <i>et al.</i> 2011)
<i>M. tuberculosis</i>	CDC1551	NC_002755.2	4, 4.1, 4.1.1, 4.1.1.3	Euro, American lineage strain, X family (Peterson <i>et al.</i> 2002)
<i>M. tuberculosis</i>	CTRI-2	CP002992.1	4, 4.3, 4.3.3	Euro, American lineage 4 strain, LAM family, LAM9 (Ilina <i>et al.</i> 2013)
<i>M. tuberculosis</i>	EAI5	CP006578.1	1, 1.1	East African Indian lineage 1, EAI5 (Rashdi & Jadhav 2014)
<i>M. tuberculosis</i>	EAI5 NITR206	NC_021194.1	1, 1.1	East African Indian lineage 1, EAI5 (Narayanan & Deshpande 2013)
<i>M. tuberculosis</i>	F11	NC_009565.1	4, 4.3, 4.3.2, 4.3.2.1	Euro, American lineage 4 strain, LAM family, LAM3 (Gagneux & Small 2007)
<i>M. tuberculosis</i>	HKBS1	CP002871.1	2, 2.2, 2.2.1	Beijing/W Lineage
<i>M. tuberculosis</i>	KZN1435	NC_012943.1	4, 4.3, 4.3.3	Euro, American lineage 4 strain, LAM family (Feuerriegel <i>et al.</i> 2010)
<i>M. tuberculosis</i>	KZN4207	NC_016768.1	4, 4.3, 4.3.3	Euro, American lineage 4 strain, LAM family (Feuerriegel <i>et al.</i> 2010)

<i>M. tuberculosis</i>	KZN605	NC_018078.1	4, 4.3, 4.3.3	Euro, American lineage 4 strain, LAM family (Feuerriegel <i>et al.</i> 2010)
<i>M. tuberculosis</i>	RGTB327	CP003233.1	4, 4.3, 4.3.4	Not reported (Madhavalatha <i>et al.</i> 2012)
<i>M. tuberculosis</i>	RGTB423	NC_017528.1	1, 1.2.2	Not reported (Madhavalatha <i>et al.</i> 2012)
<i>M. tuberculosis</i>	Strain Beijing NITR203	NC_021054.1	2, 2.2, 2.2.1, 2.2.1.1	Beijing strain (Narayanan & Deshpande 2013)
<i>M. tuberculosis</i>	Strain Erdman ATCC35801	AP012340.1	4, 4.1, 4.1.2, 4.1.2.1	Euro, American lineage 4 strain, Haarlem strain (Alix <i>et al.</i> 2006)
<i>M. tuberculosis</i>	Strain Haarlem	CP001664.1	4, 4.1, 4.1.2, 4.1.2.1	Euro, American lineage 4 strain, Haarlem strain
<i>M. tuberculosis</i>	UT205	NC016934.1	4, 4.3, 4.3.4, 4.3.4.2	Euro, American lineage 4 strain, LAM family (Isaza <i>et al.</i> 2012)

Set of 27 MTBC complete genomes downloaded from GenBank. The species, strain name and GenBank accession numbers are provided in columns 1 to 3. The predicted lineages based on the presence of strain-specific SNPs (*) in the whole genome matches the ones reported in the literature (**).

Supplementary Table 6 Lineage, specific SNPs at drug resistance genes

Drug	Gene name	Lineage	Position**	Gene Coord.	Allele Chg.	Codon Num.	Codon Chg.	AA Chg.	Locus Tag
INH	<i>katG</i>	4	2154724*	1388	C/A	463	CGG/CTG	R/L	Rv1908c
INH	<i>katG</i>	BOVAFRI	2155503	609	G/A	203	ACC/ACT	T/T	Rv1908c
INH	<i>inhA</i>	6	1674434	233	T/C	78	GTG/GCG	V/A	Rv1484
INH	<i>ahpC promoter</i>	3	2726105*	-	G/A	-	-	-	-
INH	<i>kasA</i>	4.3.3	2518919	805	G/A	269	GGT/AGT	G/S	Rv2245
INH	<i>ndh</i>	4.4.1	2102990	53	A/G	18	GTG/GCG	V/A	Rv1854c
INH	<i>ndh</i>	5	2101921	1122	C/T	374	TCG/TCA	S/S	Rv1854c
INH	<i>ndh</i>	7	2102218	825	G/A	275	GTC/GTT	V/V	Rv1854c
EMB	<i>embA</i>	1	4245969	2737	C/T	913	CCG/TCG	P/S	Rv3794
EMB	<i>embA</i>	1.1	4243848	616	G/A	206	GTG/ATG	V/M	Rv3794
EMB	<i>embA</i>	1.2.1	4244420	1188	G/C	396	GTG/GTC	V/V	Rv3794
EMB	<i>embA</i>	2.1	4246088	2856	A/G	952	CAA/CAG	Q/Q	Rv3794
EMB	<i>embA</i>	2.2	4243460*	228	C/T	76	TGC/TGT	C/C	Rv3794
EMB	<i>embA</i>	4.4.2	4246508	3276	G/A	1092	GCG/GCA	A/A	Rv3794
EMB	<i>embA</i>	4.6.1.2	4245055	1823	C/A	608	ACC/AAC	T/N	Rv3794
EMB	<i>embB</i>	2.2.1.1	4248115	1602	C/T	534	GAC/GAT	D/D	Rv3795
EMB	<i>embB</i>	4.1.1.2	4246930	417	G/C	139	CAG/CAC	Q/H	Rv3795
EMB	<i>embB</i>	4.4.1.2	4249012	2499	G/A	833	CTG/CTA	L/L	Rv3795

EMB	<i>embB</i>	4.7	4249732	3219	C/G	1073	GCC/GCG	A/A	Rv3795
EMB	<i>embB</i>	7	4248073	1560	C/T	520	ACC/ACT	T/T	Rv3795
EMB	<i>embB</i>	BOVAFRI	4246864	351	C/T	117	GTC/GTT	V/V	Rv3795
EMB	<i>embC</i>	1	4241042	1180	A/G	394	AAC/GAC	N/D	Rv3793
EMB	<i>embC</i>	3	4242075*	2213	G/A	738	CGG/CAG	R/Q	Rv3793
EMB	<i>embC</i>	3.1.1	4241562	1700	G/A	567	CGC/CAC	R/H	Rv3793
EMB	<i>embC</i>	4.1	4242803*	2941	G/C	981	GTG/CTG	V/L	Rv3793
EMB	<i>embC</i>	4.1.1.1	4240897	1035	C/G	345	CGC/CGG	R/R	Rv3793
EMB	<i>embC</i>	4.6.2.1	4242883	3021	C/T	1007	CCC/CCT	P/P	Rv3793
EMB	<i>embC</i>	4.9	4242643	2781	C/T	927	CGC/CGT	R/R	Rv3793
EMB	<i>embC</i>	7	4240153	291	G/A	97	TCG/TCA	S/S	Rv3793
EMB	<i>embR</i>	1	1417019	329	C/T	110	TGC/TAC	C/Y	Rv1267c
EMB	<i>embR</i>	4.6.1.2	1416410	938	A/C	313	CTG/CGG	L/R	Rv1267c
EMB	<i>embR</i>	4.6.2.1	1416702	646	A/G	216	TAC/CAC	Y/H	Rv1267c
EMB	<i>embR</i>	7	1416977	371	T/C	124	CAC/CGC	H/R	Rv1267c
EMB	<i>ubiA</i>	4.4.2	4268928	906	G/A	302	GGC/GGT	G/G	Rv3806c
EMB	<i>ubiA</i>	4.4.2	4269375	459	C/T	153	GTG/GTA	V/V	Rv3806c
EMB	<i>ubiA</i>	BOVAFRI	4269351	483	G/A	161	GCC/GCT	A/A	Rv3806c
EMB	<i>aftA</i>	1.2.2	4238120	189	G/A	63	CAG/CAA	Q/Q	Rv3792
EMB	<i>aftA</i>	4.1.2.1	4239298	1367	C/T	456	GCC/GTC	A/V	Rv3792
EMB	<i>aftA</i>	4.4	4238963	1032	C/T	344	CAC/CAT	H/H	Rv3792
EMB	<i>aftA</i>	5	4239843	1912	A/C	638	AAG/CAG	K/Q	Rv3792
EMB	<i>aftA</i>	7	4238778	847	G/A	283	GTG/ATG	V/M	Rv3792
EMB	<i>nuoD</i>	2.1	3513538	201	A/T	67	GAA/GAT	E/D	Rv3148
EMB	<i>nuoD</i>	4.3.4.2	3514512	1175	G/C	392	GGT/GCT	G/A	Rv3148
RMP	<i>rpoB</i>	3	762434	2628	T/G	876	GGT/GGG	G/G	Rv0667
RMP	<i>rpoB</i>	4	763031	3225	T/C	1075	GCT/GCC	A/A	Rv0667
RMP	<i>rpoC</i>	1	763884	515	C/T	172	GCC/GTC	A/V	Rv0668
RMP	<i>rpoC</i>	1	763886	517	C/A	173	CGG/AGG	R/R	Rv0668
RMP	<i>rpoC</i>	1.1	765171	1802	C/T	601	CCG/CTG	P/L	Rv0668
RMP	<i>rpoC</i>	1.1.3	765230	1861	G/A	621	GCG/ACG	A/T	Rv0668
RMP	<i>rpoC</i>	4.1	765150	1781	G/A	594	GGG/GA G	G/E	Rv0668
RMP	<i>rpoC</i>	4.3	764995	1626	C/G	542	GCC/GCG	A/A	Rv0668
RMP	<i>rpoC</i>	7	764013	644	A/C	215	GAG/GCG	E/A	Rv0668
RMP	<i>rpoC</i>	7	766955	3586	G/A	1196	GAG/AAG	E/K	Rv0668
STR	<i>rrs</i>	4.3.2	1472337	-	C/T	-	-	-	-
STR	<i>gid</i>	1	4407873	330	C/A	110	GTG/GTT	V/V	Rv3919c
STR	<i>gid</i>	1.1.3	4407780	423	C/T	141	GCG/GCA	A/A	Rv3919c
STR	<i>gid</i>	2.2	4407927	276	T/G	92	GAA/GAC	E/D	Rv3919c
STR	<i>gid</i>	4	4407588	615	T/C	205	GCA/GCG	A/A	Rv3919c
STR	<i>gid</i>	4.3	4408156	47	A/C	16	CTT/CGT	L/R	Rv3919c

FLQ	<i>gyrA</i>	1	8452	1151	C/T	384	GCA/GTA	A/V	Rv0006
FLQ	<i>gyrA</i>	1.2.1	9260	1959	G/C	653	CTG/CTC	L/L	Rv0006
FLQ	<i>gyrA</i>	3.1.2.2	9611	2310	C/T	770	GAC/GAT	D/D	Rv0006
FLQ	<i>gyrA</i>	4.3.3	8040	739	G/A	247	GGC/AGC	G/S	Rv0006
FLQ	<i>gyrA</i>	4.5	7892	591	G/A	197	CTG/CTA	L/L	Rv0006
FLQ	<i>gyrA</i>	4.6.1	7539*	238	A/G	80	ACC/GCC	T/A	Rv0006
FLQ	<i>gyrA</i>	5	9566	2265	C/T	755	TAC/TAT	Y/Y	Rv0006
FLQ	<i>gyrA</i>	7	8876	1575	C/T	525	TAC/TAT	Y/Y	Rv0006
FLQ	<i>gyrB</i>	1	6112	873	G/C	291	ATG/ATC	M/I	Rv0005
FLQ	<i>gyrB</i>	1.1.2	6124	885	C/T	295	GCC/GCT	A/A	Rv0005
FLQ	<i>gyrB</i>	4.3.2.1	5520	281	C/T	94	CCG/CTG	P/L	Rv0005
FLQ	<i>gyrB</i>	4.3.2.1	7222	1983	C/T	661	AGC/AGT	S/S	Rv0005
PZA	<i>rpsA</i>	2	1834177*	636	A/C	212	CGA/CGC	R/R	Rv1630
PZA	<i>rpsA</i>	7	1834916	1375	A/C	459	ACC/CCC	T/P	Rv1630
ETH	<i>ethA</i>	1.2.2	4326439	1035	G/T	345	AAC/AAA	N/K	Rv3854c
ETH	<i>ethA</i>	3.1.2.2	4326176	1298	T/G	433	GAG/GCG	E/A	Rv3854c
ETH	<i>ethA</i>	4.6.2.2	4326739	735	G/C	245	CGC/CGG	R/R	Rv3854c
ETH	<i>ethR</i>	4.6.2.2	4328004	456	G/A	152	GTG/GTA	V/V	Rv3855
ETH	<i>inhA</i>	6	1674434	233	T/C	78	GTG/GCG	V/A	Rv1484
AMI	<i>rrs</i>	4.3.2	1472337	-	C/T	-	-	-	-
AMI	<i>gid</i>	1	4407873*	330	C/A	110	GTG/GTT	V/V	Rv3919c
AMI	<i>gid</i>	1.1.3	4407780	423	C/T	141	GCG/GCA	A/A	Rv3919c
AMI	<i>gid</i>	2.2	4407927*	276	T/G	92	GAA/GAC	E/D	Rv3919c
AMI	<i>gid</i>	4	4407588*	615	T/C	205	GCA/GCG	A/A	Rv3919c
AMI	<i>gid</i>	4.3	4408156*	47	A/C	16	CTT/CGT	L/R	Rv3919c
CAP	<i>rrs</i>	4.3.2	1472337	-	C/T	-	-	-	-
KAN	<i>rrs</i>	4.3.2	1472337	-	C/T	-	-	-	-
AMK	<i>rrs</i>	4.3.2	1472337	-	C/T	-	-	-	-
CAP	<i>tlyA</i>	7	1918281	342	A/C	114	GGA/GGC	G/G	Rv1694

Lineage and sub-lineage specific SNPs found in DR genes. SNPs are annotated using their chromosome coordinate on the H37Rv reference NC_000962.3 (**). The alternative allele (right hand nucleotide in column four) is the one specific to the lineage or sub-lineage, with the exception of two cases (**), lineage 4 and 4.9, in which the reference allele is the one specific to the lineage. SNPs are annotated providing the locus tag, associated lineage/sub-lineage, gene name, gene coordinate, codon number, nucleotide change, codon change and resulting amino acid change. SNPs described in (Feuerriegel *et al.* 2014), a recent study reporting phylogenetic SNPs in DR genes in MTBC, are also indicated (*).

Supplementary Table 7 Non-synonymous lineage-specific SNPs at known epitopes in H37Rv

Epitope Id	Start	End	Locus Id	Amino acid sequence	Lineage_SNP
7000006195549910	157070	157129	Rv0129c	AAVGLSMSGGALILAAYYP	3_157129
7000006195549960	2952682	2952741	Rv2626c	DDRLHGMILTDRDIVIKGLA	7_2952738
7000006195549960	2952712	2952771	Rv2626c	DRDIVIKGLAAGLDPNTATA	7_2952738
7000006195549960	2955088	2955147	Rv2628	IRAVGPYAWAGRCGRIGRWG	7_2955128
7000006195549970	2955358	2955417	Rv2628	DWPAAYAIGEHLSEIIVAV	7_2955392
7000006195549970	2955058	2955117	Rv2628	MSTQRPRHSGIRAVGPYAWA	4.8_2955061
7000006195549970	2955118	2955177	Rv2628	GRCGRIGRWGVHQAEMMNLA	7_2955128
7000006195549980	1403228	1403287	Rv1255c	RRARWVVRMLTSLMFPGRD	5_1403266
7000006195550000	3079639	3079698	Rv2770c	VANALLAELTATNILGQNV	1.1.1.1_3079685
7000006195550000	3079669	3079728	Rv2770c	TATNILGQNVSAIAATEARY	1.1.1.1_3079685
7000006195550000	3079819	3079878	Rv2770c	SHITNPAGLAHQAAVQAG	4_3079877
7000006195550020	2955058	2955102	Rv2628	MSTQRPRHSGIRAVG	4.8_2955061
7000006195550020	2955088	2955132	Rv2628	IRAVGPYAWAGRCGR	7_2955128
7000006195550020	2955103	2955147	Rv2628	PYAWAGRCGRIGRWG	7_2955128
7000006195550020	2955118	2955162	Rv2628	GRCGRIGRWGVHQA	7_2955128
7000006195550040	2955358	2955402	Rv2628	DWPAAYAIGEHLSEI	7_2955392
7000006195550040	2955373	2955417	Rv2628	YAIGEHLSEIIVAV	7_2955392
7000006195549590	4352424	4352468	Rv3874	AQAAVVRFQEAANKQ	7_4352439
7000006195549590	4352439	4352483	Rv3874	VRFQEAANKQKQELD	4.1.2.1_4352475 7_4352439
7000006195549590	4352454	4352498	Rv3874	AANKQKQELDEISTN	4.1.2.1_4352475
7000006195549590	4352469	4352513	Rv3874	KQELDEISTNIRQAG	4.1.2.1_4352475
7000006195549600	686965	687012	Rv0589	VAFRAGLVMEAGSKVT	4.9_686972
7000006195549610	4351141	4351194	Rv3873	PMLAAAAGWQTLAALDA	4.1.2.1_4351160
7000006195549620	4351723	4351776	Rv3873	GPMQQLTQPLQQVTSLS	1_4351759
7000006195549620	4351753	4351806	Rv3873	QQVTSLSFQVGGTGGG	1_4351759
7000006195549640	352028	352081	Rv0288	AMEDLVRAYHAMSSTHEA	6_352058
7000006195549640	352058	352111	Rv0288	AMSSTHEANTMAMMARDT	6_352058
7000006195549640	3378771	3378830	Rv3019c	YAGTLQSLGADIASEQAVLS	1_3378828
7000006195549640	3378801	3378860	Rv3019c	DIASEQAVLSSAWQDGTGIT	1_3378828
7000006195549650	3378921	3378980	Rv3019c	SMSGTHESNTMAMLARDGAE	7_3378952
7000006195549650	3378951	3378998	Rv3019c	MAMLARDGAEAAKWGG	7_3378952
7000006195549670	4352418	4352462	Rv3874	TAAQAAVVRFQEAAN	7_4352439
7000006195549670	4352430	4352474	Rv3874	AAVRFQEAANKQKQ	7_4352439
7000006195549670	4352442	4352486	Rv3874	RFQEAANKQKQELDE	4.1.2.1_4352475
7000006195549670	4352466	4352510	Rv3874	QKQELDEISTNIRQA	4.1.2.1_4352475
7000006195549690	4352424	4352483	Rv3874	AQAAVVRFQEAANKQKQELD	4.1.2.1_4352475 7_4352439
7000006195549690	4352454	4352513	Rv3874	AANKQKQELDEISTNIRQAG	4.1.2.1_4352475

7000006195549730	1020433	1020477	Rv0915c	LGQNSAAIAATQAEY	5_1020452
7000006195549780	4352409	4352483	Rv3874	AAGTAAQAAVVRFQEAANKQKQELD	4.1.2.1_4352475 7_4352439
7000006195549780	4352454	4352528	Rv3874	AANKQKQELDEISTNIRQAGVQYSR	4.1.2.1_4352475
7000006195549790	3187480	3187554	Rv2875	GASVTVTGGGNSLKVGNADVCCGV	4.3.1_3187535
7000006195549790	3187525	3187599	Rv2875	GNADVCCGVSTANATVYMIDSVLM	4.3.1_3187535
7000006195549810	4357092	4357151	Rv3878	AAELAPRVVATVPQLVQLAP	4.1_4357123
7000006195549820	4266002	4266061	Rv3804c	FYSDWYQPACGKAGCQTYKW	4.4.1.2_4266036
7000006195549820	4266032	4266091	Rv3804c	GKAGCQTYKWETFLTSELPG	4.4.1.2_4266036
7000006195549850	4265981	4266040	Rv3804c	PVGGQSSFYSDWYQPACGKA	4.4.1.2_4266036
7000006195549850	4266011	4266070	Rv3804c	DWYQPACGKAGCQTYKWETF	4.4.1.2_4266036
7000006195549570	2227288	2227359	Rv1983	NGIVTAPTAVNVVLLSIPTSPFAI	7_2227339
7000006195549570	3351437	3351508	Rv2994	PSWGLVVTMFAWGYLLDHVGERMV	1_3351472
7000006195549580	4357107	4357175	Rv3878	PRVVATVPQLVQLAPHAVQMSQN	4.1_4357123

List of epitopes extracted from *Immune Epitope Database* (www.iedb.org) containing lineage specific SNPs. The last column indicates the lineage and SNP chromosome coordinate found within each epitope.

Supplementary Table 8 Sub-lineage proportions observed in the Russian dataset (n=850) (Casali *et al.* 2014) when using the 62-SNPs classification scheme.

Sub-lineage	Frequency
2.2.1 (Modern Beijing)	518
4.8 (T spoligotype, RD219)	71
4.2.1 (Ural)	68
4.3.3 (LAM, RD115)	57
4.1.2.1 (Haarlem, RD182)	36
4.1.2	15
4.1 (X, type)	8
3 (CAS, Delhi)	5
4.9 (H37Rv, like)	5
BOV (<i>M. bovis</i>)	4
4.3.4.1 (LAM, RD174)	3
4.3.4.2 (LAM, RD174)	3
4.4.1.1 (S type)	3
4.7	3
5 (West, Africa 1)	2
6 (West, Africa 2)	2
1.1.2 (EAI5, EAI3)	1
1.1.3 (EAI6)	1
2.2.1.1 (Beijing, RD150)	1
4.5 (RD122)	1
Probable mixed infections*	41

*The most frequent combinations of strain types were: Modern Beijing (2.2.1) with LAM (4.3.3) (n=8), Modern Beijing with Ural (4.2.1) (n=8), Modern Beijing with Haarlem (4.1.2.1) (n=4), Modern Beijing with 4.8 (n=3), and Ural with 4.8.

Supplementary Table 9 Probable cases of Karonga-Malawian mixed samples

Isolate accession number	<i>SpolPred</i> -derived octal code	<i>Spolpred</i> -derived spoligotype	Median DOC	Predicted lineages and sub-lineages
ERR036233	47777777673771	Orphan	82	1, 1.1, 1.1.2, 4.3, 4.3.4, 4.3.4.1
ERR036248	77777777700371	T	108	4, 4.3.4, 4.3.4.2.1, 4.8
ERR037469	77777777473771	Manu1	351	1, 1.2.2, 4.3, 4.3.4, 4.3.4.2, 4.3.4.2.1
ERR161049	777777606060771	LAM11-ZWE	69	2.2.1, 4, 4.3, 4.3.4, 4.3.4.2, 4.3.4.2.1
ERR161050	000000200003771	Orphan	53	2, 2.2, 2.2.1, 4.3, 4.3.4, 4.3.4.2, 4.3.4.2.1
ERR161071	713777753003771	Orphan	154	3, BOV-AFRI
ERR161077	721777746413771	Orphan	72	1, 1.1, 1.1.3, 4.1, 4.1.1, 4.1.1.3
ERR161078	7737777773771	Orphan	69	1, 1.1, 1.1.3, 4.1, 4.1.1, 4.1.1.3
ERR161123	757777737450031	Orphan	89	1, 1.2.2, 4.3, 4.3.4.2, 4.3.4.2.1
ERR163947	40407777413771	Orphan	57	1, 1.1, 1.1.2, 1.2.2
ERR164021	77777777473771	Manu1	188	1, 1.2.2, 4.3, 4.3.4, 4.3.4.2, 4.3.4.2.1
ERR176549	000000004060731	Orphan	62	4, 4.2, 4.3, 4.3.4, 4.3.4.2, 4.3.4.2.1, 4.6.1, 4.6.2.1, 4.6.2.2
ERR176611	000000000003771	Beijing	68	2, 2.2, 2.2.1, 4.3.4.2
ERR176616	777777606060771	LAM11-ZWE	60	1, 1.1, 1.1.3, 4.3, 4.3.4, 4.3.4.2, 4.3.4.2.1
ERR176620	777764207360771	Orphan	60	4, 4.3, 4.3.3, 4.3.4, 4.3.4.2, 4.3.4.2.1
ERR176652	700777746003371	Orphan	40	1, 1.1, 1.1.3, 4.3.4
ERR176653	700076777360771	Orphan	50	4, 4.1, 4.1.1, 4.1.1.3, 4.3, 4.3.4.2.1
ERR176661	47747777410571	Orphan	47	1, 1.1, 1.1.2, 4.6
ERR176709	503377400041771	Orphan	53	3, 3.1.1, 4.3, 4.3.4, 4.3.4.2, 4.3.4.2.1
ERR181686	777777606060771	LAM11-ZWE	79	3.1.1, 4.3, 4.3.4, 4.3.4.2, 4.3.4.2.1
ERR181705	777777616462671	Orphan	108	1, 1.2.2, 4, 4.3, 4.3.4, 4.3.4.2, 4.3.4.2.1
ERR181782	703377404001771	Orphan	130	3, 3.1.1, 4.1.1
ERR181811	700777747433771	Orphan	52	1, 1.1, 1.1.3, 4.1, 4.1.2, 4.1.2.1
ERR181813	777777706473771	Orphan	66	1, 1.1, 1.1.3, 4.3, 4.3.4, 4.3.4.2, 4.3.4.2.1
ERR181974	777777606060771	LAM11-ZWE	66	4, 4.3, 4.3.4, 4.3.4.2, 4.3.4.2.1, 4.6.1.1
ERR181977	477740017453731	Orphan	90	1, 1.1, 1.1.2, 4.1.2
ERR182003	000000007760771	Orphan	89	4, 4.3, 4.3.3, 4.9
ERR182015	77777775760731	T2	71	4, 4.3, 4.6, 4.6.1, 4.6.1.2
ERR182026	720777746013771	Orphan	72	1, 1.1, 1.1.3, 4.1
ERR182027	77773777760731	AmbiguousT3-T2	82	4, 4.1, 4.1.2, 4.3, 4.3.4.2
ERR182041	63777477760730	T2-Uganda	73	4, 4.3.4.2, 4.6, 4.6.1, 4.6.1.1
ERR190343	703777740003771	CAS1-Delhi	93	3, 4.3, 4.3.4, 4.3.4.2.1
ERR190379	577761377410771	Orphan	89	1, 1.1, 1.1.2, 4.3, 4.3.4, 4.3.4.2, 4.3.4.2.1

ERR212005	727377404001771	Orphan	59	3, 3.1.1, 4.3, 4.3.4
ERR212098	000036500003771	Orphan	134	1, 2, 2.2, 2.2.1
ERR216914	777753777760671	Orphan	50	4, 4.3, 4.3.4, 4.3.4.2, 4.3.4.2.1, 4.4, 4.4.1, 4.4.1.2
ERR221561	77777607760731	LAM4	58	4, 4.3, 4.3.3, 4.3.4, 4.3.4.2
ERR221567	70077747413771	EAI6-BGD1	63	1, 1.1, 1.1.3, 4.3.4.2.1
ERR245754	77777777763771	Manu2	200	2, 2.2, 2.2.1, 4.1, 4.1.2
ERR245795	777377407761771	Orphan	132	3, 3.1.1, 4.3, 4.3.2, 4.3.2.1
ERR245797	67777607760771	LAM1	175	4, 4.3, 4.3.4, 4.3.4.1, 4.3.4.2, 4.3.4.2.1

Possible cases of mixed samples, i.e. harbouring lineage-specific SNPs from multiple sub-lineages or lineages.

Supplementary Table 10 Locus-based GWAS top hits

Drug/Model	Locus	Gene name	P-value	Functional category	Function/Product
INH/A	Rv1908c	katG	2.91E-115	virulence, detoxification, adaptation	Catalase-peroxidase-peroxynitritase T KatG
	Rv0667	rpoB	4.56E-108	information pathways	DNA-directed RNA polymerase (beta chain) RpoB (transcriptase beta chain) (RNA polymerase beta subunit)
	Rv3795	embB	4.27E-81	cell wall and cell processes	Integral membrane indolylacetyltransferase EmbB
	Rv2043c	pncA	3.30E-56	intermediary metabolism and respiration	Pyrazinamidase/nicotinamidase PncA (PZase)
	Rv1482c-Rv1483	fabG1 promoter	2.78E-48	lipid metabolism	3-oxoacyl-[acyl-carrier protein] reductase FabG1 (3-ketoacyl-acyl carrier protein reductase) (mycolic acid biosynthesis a protein)
	Rv0682	rpsL	2.66E-40	information pathways	30S ribosomal protein S12 RpsL
	Rv3919c	gid	2.55E-39	cell wall and cell processes	Probable glucose-inhibited division protein B Gid
	Rv1816	-	2.91E-34	regulatory proteins	Possible transcriptional regulatory protein
	Rv1484	inhA	2.54E-32	lipid metabolism	NADH-dependent enoyl-[acyl-carrier-protein] reductase InhA (NADH-dependent enoyl-ACP reductase)
	Rv3854c	ethA	3.57E-30	intermediary metabolism and respiration	Monooxygenase EthA

	rrs	rrs	7.08E-29	stable RNAs	Ribosomal RNA 16S
	Rv3069	-	6.15E-27	cell wall and cell processes	Probable conserved transmembrane protein
	Rv3746c	PE34	1.32E-26	PE/PPE	Probable PE family protein PE34 (PE family-related protein)
	Rv2202c-Rv2203	Rv2203 promoter	1.47E-25	cell wall and cell processes	Possible conserved membrane protein
	Rv3664c	dppC	5.92E-24	cell wall and cell processes	Probable dipeptide-transport integral membrane protein ABC transporter DppC
INH/B	Rv1908c	katG	6.28E-47	virulence, detoxification, adaptation	Catalase-peroxidase-peroxynitritase T KatG
	Rv1482c-Rv1483	fabG1 promoter	5.43E-22	lipid metabolism	3-oxoacyl-[acyl-carrier protein] reductase FabG1 (3-ketoacyl-acyl carrier protein reductase) (mycolic acid biosynthesis a protein)
	Rv0671	lpqP	4.77E-14	cell wall and cell processes	Possible conserved lipoprotein LpqP
	Rv3746c	PE34	7.76E-12	PE/PPE	Probable PE family protein PE34 (PE family-related protein)
	Rv0275c	-	8.42E-12	regulatory proteins	Possible transcriptional regulatory protein (possibly TetR-family)
	Rv3068c	pgmA	2.61E-11	intermediary metabolism and respiration	Probable phosphoglucomutase PgmA (glucose phosphomutase) (PGM)
	Rv0667	rpoB	9.26E-11	information pathways	DNA-directed RNA polymerase (beta chain) RpoB (transcriptase beta chain) (RNA polymerase beta subunit)
Rv3919c	gid	9.83E-11	cell wall and cell processes	Probable glucose-inhibited division protein B Gid	
RMP/A	Rv0667	rpoB	4.97E-120	information pathways	DNA-directed RNA polymerase (beta chain) RpoB (transcriptase beta chain) (RNA polymerase beta subunit)
	Rv1908c	katG	1.24E-102	virulence, detoxification, adaptation	Catalase-peroxidase-peroxynitritase T KatG
	Rv3795	embB	1.30E-88	cell wall and cell processes	Integral membrane indolyacetylinositol arabinosyltransferase EmbB (arabinosylindolyacetylinositol synthase)
	Rv2043c	pncA	3.32E-73	intermediary metabolism and respiration	Pyrazinamidase/nicotinamidase PncA (PZase)
	Rv1482c-Rv1483	fabG1 promoter	3.61E-51	lipid metabolism	3-oxoacyl-[acyl-carrier protein] reductase FabG1 (3-ketoacyl-acyl carrier protein reductase) (mycolic acid biosynthesis a protein)

	Rv0682	rpsL	5.83E-51	information pathways	30S ribosomal protein S12 RpsL
	Rv1484	inhA	4.40E-40	lipid metabolism	NADH-dependent enoyl-[acyl-carrier-protein] reductase InhA (NADH-dependent enoyl-ACP reductase)
	Rv1816	-	5.67E-31	regulatory proteins	Possible transcriptional regulatory protein
	Rv0668	rpoC	1.63E-30	information pathways	DNA-directed RNA polymerase (beta' chain) RpoC (transcriptase beta' chain) (RNA polymerase beta' subunit).
	Rv3919c	gid	5.40E-29	cell wall and cell processes	Probable glucose-inhibited division protein B Gid
	Rv2904c- Rv2905	lppW promoter	1.08E-26	cell wall and cell processes	Probable conserved alanine rich lipoprotein LppW
	rrs	rrs	2.04E-26	stable RNAs	Ribosomal RNA 16S
	Rv2202c- Rv2203	Rv2203 promoter	2.20E-26	cell wall and cell processes	Possible conserved membrane protein
	Rv3854c	ethA	2.60E-26	intermediary metabolism and respiration	Monooxygenase EthA
	Rv2000	-	9.14E-26	conserved hypotheticals	Unknown protein
	Rv3069	-	1.50E-25	cell wall and cell processes	Probable conserved transmembrane protein
	Rv0751c	mmsB	1.77E-25	intermediary metabolism and respiration	Probable 3-hydroxyisobutyrate dehydrogenase MmsB (hibadh)
	Rv3792	aftA	1.82E-25	cell wall and cell processes	Arabinofuranosyltransferase AftA
RMP/B	Rv0667	rpoB	2.22E-05	information pathways	DNA-directed RNA polymerase (beta chain) RpoB (transcriptase beta chain) (RNA polymerase beta subunit)
	Rv3795	embB	3.72E-25	cell wall and cell processes	Integral membrane indolylacetylinsitol arabinosyltransferase EmbB (arabinosylindolylacetylinsitol synthase)
	Rv2043c	pncA	2.33E-24	intermediary metabolism and respiration	Pyrazinamidase/nicotinamidase PncA (PZase)
EMB/A	Rv0667	rpoB	1.63E-16	information pathways	DNA-directed RNA polymerase (beta chain) RpoB (transcriptase beta chain) (RNA polymerase beta subunit)
	Rv3793- Rv3794	embA promoter	1.36E-12	cell wall and cell processes	Integral membrane indolylacetylinsitol arabinosyltransferase EmbA (arabinosylindolylacetylinsitol synthase)
	Rv1908c	katG	7.15E-11	virulence, detoxification,	Catalase-peroxidase-peroxynitritase T KatG

				adaptation	
	Rv0187	-	1.33E-06	intermediary metabolism and respiration	Probable O-methyltransferase
	Rv0870c-Rv0871	fadE10 promoter	4.15E-06	lipid metabolism	Probable acyl-CoA dehydrogenase FadE10
	Rv1316c	ogt	5.23E-06	information pathways	Methylated-DNA--protein-cysteine methyltransferase Ogt (6-O-methylguanine-DNA methyltransferase) (O-6-methylguanine-DNA-alkyltransferase)
	Rv0819	mshD	7.92E-06	intermediary metabolism and respiration	GCN5-related N-acetyltransferase, MshD
	Rv0365c	-	9.55E-06	conserved hypotheticals	Conserved protein
	Rv3795	embB	9.90E-05	cell wall and cell processes	Integral membrane indolylacetylaminositol arabinosyltransferase EmbB
	Rv3793-Rv3794	embA promoter	0.001265	cell wall and cell processes	Integral membrane indolylacetylaminositol arabinosyltransferase EmbA
	Rv0035-Rv0036c	-	0.002221	-	-
	Rv2043c	pncA	0.00248	intermediary metabolism and respiration	Pyrazinamidase/nicotinamidase PncA (PZase)
EMB/B	Rv1452c-Rv1453	PE PGRS28 promoter	0.004056	cell wall and cell processes	PE-PGRS family protein PE PGRS28
	Rv1668c	-	0.008609	cell wall and cell processes	Probable first part of macrolide-transport ATP-binding protein ABC transporter
	Rv0061c	-	0.008997	conserved hypotheticals	Hypothetical protein
	Rv1418-Rv1419	lprH promoter	0.009261	cell wall and cell processes	Probable lipoprotein LprH
	Rv2043c	pncA	8.15E-36	intermediary metabolism and respiration	Pyrazinamidase/nicotinamidase PncA (PZase)
	Rv0667	rpoB	2.82E-35	information pathways	DNA-directed RNA polymerase (beta chain) RpoB (transcriptase beta chain) (RNA polymerase beta subunit)
PZA/A	Rv3795	embB	9.23E-30	cell wall and cell processes	Integral membrane indolylacetylaminositol arabinosyltransferase EmbB
	Rv1179c-Rv1180	papA3 promoter	1.97E-19	lipid metabolism	Probable conserved polyketide synthase associated protein PapA3
	Rv1482c-Rv1483	fabG1 promoter	1.78E-18	lipid metabolism	3-oxoacyl-[acyl-carrier protein] reductase FabG1 (3-ketoacyl-acyl carrier protein reductase) (mycolic acid biosynthesis a

	Rv1908c	katG	1.19E-17	virulence, detoxification, adaptation information pathways	Catalase-peroxidase- peroxynitritase T KatG
	Rv0682	rpsL	1.55E-16	information pathways	30S ribosomal protein S12 RpsL
	Rv1484	inhA	1.92E-16	lipid metabolism	NADH-dependent enoyl-[acyl-carrier-protein] reductase InhA (NADH-dependent enoyl-ACP reductase)
	Rv3475- Rv3476c	-	4.51E-15	-	-
	Rv1816	-	1.57E-14	regulatory proteins	Possible transcriptional regulatory protein
	Rv1452c	PE PGRS28	1.59E-14	PE/PPE	PE-PGRS family protein PE PGRS28
	Rv0143c	-	1.92E-14	cell wall and cell processes	Probable conserved transmembrane protein
	Rv2043c	pncA	5.41E-05	intermediary metabolism and respiration	Pyrazinamidase/nicotinamidase PncA (PZase)
	Rv3347c	PPE55	0.000495	PE/PPE	PPE family protein PPE55
	Rv1091	PE PGRS22	0.001347	PE/PPE	PE-PGRS family protein PE PGRS22
	Rv0980c	PE PGRS18	0.001424	PE/PPE	PE-PGRS family protein PE PGRS18
PZA/B	Rv2294- Rv2295	Rv2295 promoter	0.00156	Conserved hypothetical	Conserved hypothetical protein
	Rv0578c	PE PGRS7	0.001877	PE/PPE	PE-PGRS family protein PE PGRS7
	Rv2736c	recX	0.004461	information pathways	Regulatory protein RecX
	Rv2490c	PE PGRS43	0.004543	PE/PPE	PE-PGRS family protein PE PGRS43
	Rv0104	-	0.007272	conserved hypotheticals	Conserved hypothetical protein
	Rv2059	-	0.008331	conserved hypotheticals	Conserved hypothetical protein
	Rv0682	rpsL	4.68E-37	information pathways	30S ribosomal protein S12 RpsL
	Rv0667	rpoB	2.30E-33	information pathways	DNA-directed RNA polymerase (beta chain) RpoB (transcriptase beta chain) (RNA polymerase beta subunit)
STR/A	Rv1484	inhA	1.18E-26	lipid metabolism	NADH-dependent enoyl-[acyl-carrier-protein] reductase InhA (NADH-dependent enoyl-ACP reductase)
	Rv1482c- Rv1483	fabG1 promoter	2.36E-26	lipid metabolism	3-oxoacyl-[acyl-carrier protein] reductase FabG1 (3-ketoacyl-acyl carrier protein reductase) (mycolic acid biosynthesis a protein)
	Rv2202c-	Rv2203	2.77E-25	cell wall and cell	Possible conserved membrane

	Rv2203	promoter		processes	protein
	Rv3795	embB	1.17E-24	cell wall and cell processes	Integral membrane indolylacetylinoitol arabinosyltransferase EmbB (arabinosylindolylacetylinoitol synthase)
	Rv0143c	-	2.87E-23	cell wall and cell processes	Probable conserved transmembrane protein
	Rv1816	-	3.16E-22	regulatory proteins	Possible transcriptional regulatory protein
	Rv3792	aftA	3.44E-20	cell wall and cell processes	Arabinofuranosyltransferase AftA
	Rv1908c	katG	1.09E-19	virulence, detoxification, adaptation	Catalase-peroxidase-peroxynitritase T KatG
	Rv3069	-	4.50E-19	cell wall and cell processes	Probable conserved transmembrane protein
	Rv0211	pckA	4.75E-19	intermediary metabolism and respiration	Probable iron-regulated phosphoenolpyruvate carboxykinase [GTP] PckA (phosphoenolpyruvate carboxylase) (PEPCK)(pep carboxykinase)
	Rv1029	kdpA	9.81E-19	cell wall and cell processes	Probable potassium-transporting ATPase a chain KdpA (potassium-translocating ATPase a chain) (ATP phosphohydrolase [potassium-transporting] a chain) (potassium binding and translocating subunit A)
	Rv1148c	-	1.82E-18	insertion seqs and phages	Conserved hypothetical protein
	Rv0682	rpsL	2.05E-07	information pathways	30S ribosomal protein S12 RpsL
	Rv2368c	phoH1	6.24E-06	intermediary metabolism and respiration	Probable PHOH-like protein PhoH1 (phosphate starvation-inducible protein PSIH)
	Rv3318	sdhA	1.32E-05	intermediary metabolism and respiration	Probable succinate dehydrogenase
STR/B	Rv0932c-Rv0933	pstB promoter	2.06E-05	cell wall and cell processes	Phosphate-transport ATP-binding protein ABC transporter PstB
	Rv3822	-	2.15E-05	conserved hypotheticals	Conserved hypothetical protein
	Rv0147	-	2.50E-05	intermediary metabolism and respiration	Probable aldehyde dehydrogenase (NAD+) dependent
	Rv3132c	devS	2.97E-05	regulatory proteins	Two component sensor histidine kinase DevS
	Rv0303	-	3.16E-05	intermediary metabolism and respiration	Probable dehydrogenase/reductase

ETH/A	Rv1482c- Rv1483	fabG1 promoter	5.30E-43	lipid metabolism	3-oxoacyl-[acyl-carrier protein] reductase FabG1 (3-ketoacyl- acyl carrier protein reductase) (mycolic acid biosynthesis a protein)
	Rv1816	-	9.05E-35	regulatory proteins	Possible transcriptional regulatory protein
	Rv1484	inhA	3.10E-33	lipid metabolism	NADH-dependent enoyl-[acyl- carrier-protein] reductase InhA (NADH-dependent enoyl-ACP reductase)
	Rv0667	rpoB	5.94E-32	information pathways	DNA-directed RNA polymerase (beta chain) RpoB (transcriptase beta chain) (RNA polymerase beta subunit)
	Rv3795	embB	6.19E-29	cell wall and cell processes	Integral membrane indolylacetylinositol arabinylosyltransferase EmbB (arabinylosylindolylacetylinositol synthase)
	Rv2202c- Rv2203	Rv2203 promoter	7.37E-26	cell wall and cell processes	Possible conserved membrane protein
	Rv0751c	mmsB	1.68E-25	intermediary metabolism and respiration	Probable 3-hydroxyisobutyrate dehydrogenase MmsB (hibadh)
	Rv0682	rpsL	1.95E-25	information pathways	30S ribosomal protein S12 RpsL
	Rv0143c	-	4.58E-25	cell wall and cell processes	Probable conserved transmembrane protein
	Rv2000	-	6.80E-25	conserved hypotheticals	Unknown protein
ETH/B	Rv1816	-	0.000557	regulatory proteins	Possible transcriptional regulatory protein
	Rv1484	inhA	0.000587	lipid metabolism	NADH-dependent enoyl-[acyl- carrier-protein] reductase InhA (NADH-dependent enoyl-ACP reductase)
	Rv0194	-	0.000782	cell wall and cell processes	Probable transmembrane multidrug efflux pump
	Rv1482c- Rv1483	fabG1 promoter	0.001545	lipid metabolism	3-oxoacyl-[acyl-carrier protein] reductase FabG1 (3-ketoacyl- acyl carrier protein reductase) (mycolic acid biosynthesis a protein)
	Rv0037c	-	0.008639	cell wall and cell processes	Probable conserved integral membrane protein
	Rv3148	nuoD	0.008639	intermediary metabolism and respiration	Probable NADH dehydrogenase I (chain D) NuoD (NADH- ubiquinone oxidoreductase chain D)
	Rv1872c- Rv1873	-	0.01181	-	-
	Rv2124c	methH	0.012309	intermediary metabolism and	5-methyltetrahydrofolate-- homocystein methyltransferase

				respiration	MethH (methionine synthase, vitamin-B12 dependent isozyme) (ms)
	Rv3897c	-	0.012757	conserved hypotheticals	Conserved hypothetical protein
	Rv0946c	pgi	0.013046	intermediary metabolism and respiration	Probable glucose-6-phosphate isomerase Pgi (GPI)
OFX/A	Rv0006	gyrA	2.55E-12	information pathways	DNA gyrase (subunit A) GyrA (DNA topoisomerase (ATP-hydrolysing)) (DNA topoisomerase II) (type II DNA topoisomerase)
	Rv3795	embB	1.48E-11	cell wall and cell processes	Integral membrane indolylacetylinsitol arabinosyltransferase EmbB (arabinosylindolylacetylinsitol synthase)
	Rv2043c	pncA	8.36E-10	intermediary metabolism and respiration	Pyrazinamidase/nicotinamidase PncA (PZase)
	Rv3793-Rv3794	embA promoter	6.87E-09	cell wall and cell processes	Integral membrane indolylacetylinsitol arabinosyltransferase EmbA (arabinosylindolylacetylinsitol synthase)
	Rv1482c-Rv1483	fadG1 promoter	9.12E-08	lipid metabolism	3-oxoacyl-[acyl-carrier protein] reductase FabG1 (3-ketoacyl-acyl carrier protein reductase) (mycolic acid biosynthesis a protein)
	Rv2416c-Rv2417c	eis promoter	1.39E-07	virulence, detoxification, adaptation	Enhanced intracellular survival protein Eis, GCN5-related N-acetyltransferase
	Rv1484	inhA	3.01E-07	lipid metabolism	NADH-dependent enoyl-[acyl-carrier-protein] reductase InhA (NADH-dependent enoyl-ACP reductase)
	Rv3423c	alr	6.93E-07	intermediary metabolism and respiration	Alanine racemase Alr
	Rv3069	-	3.48E-05	cell wall and cell processes	Probable conserved transmembrane protein
	Rv1816	-	4.21E-05	regulatory proteins	Possible transcriptional regulatory protein
OFX/B	Rv0006	gyrA	0.012655	information pathways	DNA gyrase (subunit A) GyrA (DNA topoisomerase (ATP-hydrolysing)) (DNA topoisomerase II) (type II DNA topoisomerase)
	Rv1663	pks17	0.021396	lipid metabolism	Probable polyketide synthase Pks17
	Rv2799	-	0.021931	cell wall and cell processes	Probable membrane protein

	Rv0206c	mmpL3	0.026943	cell wall and cell processes	Possible conserved transmembrane transport protein MmpL3
	Rv3425	PPE57	0.027406	PE/PPE	PPE family protein PPE57
	Rv2416c- Rv2417c	eis promoter	0.028844	virulence, detoxification, adaptation	Enhanced intracellular survival protein Eis, GCN5-related N-acetyltransferase
	Rv0620	galK	0.034167	intermediary metabolism and respiration	Probable galactokinase GalK (galactose kinase)
	Rv1830	-	0.035624	conserved hypotheticals	Conserved hypothetical protein
	Rv3428c	-	0.038173	insertion seqs and phages	Possible transposase
	Rv3090	-	0.042676	conserved hypotheticals	Unknown alanine and valine rich protein
	rrs	rrs	7.65E-17	stable RNAs	Ribosomal RNA 16S
	Rv3795	embB	7.45E-14	cell wall and cell processes	Integral membrane indolylacetylinoitol arabinosyltransferase EmbB (arabinosylindolylacetylinoitol synthase)
	Rv2181	-	1.60E-11	cell wall and cell processes	Alpha(1->2)mannosyltransferase
	Rv3736	-	2.47E-11	regulatory proteins	Transcriptional regulatory protein (probably AraC/XylS-family)
AMK/A	Rv3378c	-	5.53E-11	intermediary metabolism and respiration	Diterpene synthase
	Rv2404c	lepA	6.09E-11	intermediary metabolism and respiration	Probable GTP-binding protein LepA (GTP-binding elongation factor)
	Rv2195	qcrA	8.98E-11	intermediary metabolism and respiration	Probable rieske iron-sulfur protein QcrA
	Rv1698	mctB	1.72E-10	cell wall and cell processes	Outer membrane protein MctB
	Rv1311- Rv1312	Rv1312 promoter	2.07E-10	cell wall and cell processes	Conserved hypothetical secreted protein
	Rv2075c- Rv2076c	Rv2075c promoter	2.07E-10	cell wall and cell processes	Possible hypothetical exported or envelope protein
	rrs	rrs	5.77E-09	stable RNAs	Ribosomal RNA 16S
	Rv3378c	-	6.67E-06	intermediary metabolism and respiration	Diterpene synthase
AMK/B	Rv3795	embB	2.00E-05	cell wall and cell processes	Integral membrane indolylacetylinoitol arabinosyltransferase EmbB (arabinosylindolylacetylinoitol synthase)
	Rv2181	-	2.44E-05	cell wall and cell processes	Alpha(1->2)mannosyltransferase
	Rv3736	-	2.59E-05	regulatory	Transcriptional regulatory

	Rv2404c	lepA	5.60E-05	intermediary metabolism and respiration	protein (probably AraC/XylS-family) Probable GTP-binding protein LepA (GTP-binding elongation factor)
	Rv2190c	-	6.02E-05	virulence, detoxification, adaptation	Conserved hypothetical protein
	Rv0337c	aspC	6.18E-05	intermediary metabolism and respiration	Probable aspartate aminotransferase AspC (transaminase A) (ASPAT) Probable Sn-glycerol-3-phosphate transport ATP-binding protein ABC transporter UgpC
	Rv2832c	ugpC	7.18E-05	cell wall and cell processes	Probable riboflavin biosynthesis protein RibA1 (GTP cyclohydrolase II)
	Rv1940	ribA1	8.49E-05	intermediary metabolism and respiration	
	Rv3795	embB	1.65E-12	cell wall and cell processes	Integral membrane indolylacetyl-inositol arabinosyltransferase EmbB
	rrs	rrs	3.33E-11	stable RNAs	Ribosomal RNA 16S
	Rv3793-Rv3794	embA promoter	4.30E-11	cell wall and cell processes	Integral membrane indolylacetyl-inositol arabinosyltransferase EmbA
	Rv1482c-Rv1483	fabG1 promoter	5.26E-10	lipid metabolism	3-oxoacyl-[acyl-carrier protein] reductase FabG1 (3-ketoacyl-acyl carrier protein reductase) (mycolic acid biosynthesis a protein)
CAP/A	Rv2195	qcrA	1.14E-09	intermediary metabolism and respiration	Probable rieske iron-sulfur protein QcrA
	Rv3736	-	2.41E-09	regulatory proteins	Transcriptional regulatory protein (probably AraC/XylS-family)
	Rv0013	trpG	2.06E-08	intermediary metabolism and respiration	Possible anthranilate synthase component II TrpG (glutamine amidotransferase)
	Rv2312-Rv2313c	-	2.12E-08	-	-
	Rv1940	ribA1	2.22E-08	intermediary metabolism and respiration	Probable riboflavin biosynthesis protein RibA1 (GTP cyclohydrolase II)
	Rv2181	NA	2.34E-08	cell wall and cell processes	Alpha(1->2)mannosyltransferase
	rrs	rrs	7.26E-07	stable RNAs	Ribosomal RNA 16S
CAP/B	Rv1482c-Rv1483	fabG1 promoter	2.15E-06	lipid metabolism	3-oxoacyl-[acyl-carrier protein] reductase FabG1 (3-ketoacyl-acyl carrier protein reductase) (mycolic acid biosynthesis a protein)
	Rv3793-Rv3794	embA promoter	2.57E-06	cell wall and cell processes	Integral membrane indolylacetyl-inositol

					arabinosyltransferase EmbA
	Rv2195	qcrA	2.29E-05	intermediary metabolism and respiration	Probable rieske iron-sulfur protein QcrA
	Rv3736	NA	3.85E-05	regulatory proteins	Transcriptional regulatory protein (probably AraC/XylS-family)
	Rv1940	ribA1	0.000119	intermediary metabolism and respiration	Probable riboflavin biosynthesis protein RibA1 (GTP cyclohydrolase II)
	Rv0013	trpG	0.000137	intermediary metabolism and respiration	Possible anthranilate synthase component II TrpG (glutamine amidotransferase)
	Rv1816	-	0.000171	regulatory proteins	Possible transcriptional regulatory protein
	Rv0235c	-	0.00018	cell wall and cell processes	Probable conserved transmembrane protein
	Rv1230c	-	0.000188	cell wall and cell processes	Possible membrane protein
	Rv3148	nuoD	0.000258	intermediary metabolism and respiration	Probable NADH dehydrogenase I (chain D) NuoD (NADH-ubiquinone oxidoreductase chain D)
	Rv3795	embB	0.00029	cell wall and cell processes	Integral membrane indolylacetylinsitol arabinosyltransferase EmbB
KAN/A	rrs	rrs	5.97E-13	stable RNAs	Ribosomal RNA 16S
	Rv2416c-Rv2417c	eis promoter	6.65E-12	virulence, detoxification, adaptation	Enhanced intracellular survival protein Eis, GCN5-related N-acetyltransferase
	Rv3795	embB	1.12E-11	cell wall and cell processes	Integral membrane indolylacetylinsitol arabinosyltransferase EmbB
	Rv2043c	pncA	2.30E-09	intermediary metabolism and respiration	Pyrazinamidase/nicotinamidase PncA (PZase)
	Rv3793-Rv3794	embA promoter	1.03E-07	cell wall and cell processes	Integral membrane indolylacetylinsitol arabinosyltransferase EmbA
KAN/B	rrs	rrs	1.24E-08	stable RNAs	Ribosomal RNA 16S
	Rv2416c-Rv2417c	eis promoter	1.51E-07	virulence, detoxification, adaptation	Enhanced intracellular survival protein Eis, GCN5-related N-acetyltransferase
	Rv3067	-	9.19E-06	conserved hypotheticals	Conserved hypothetical protein
	Rv3899c	-	9.75E-06	conserved hypotheticals	Conserved hypothetical protein
	Rv3189	-	1.59E-05	conserved hypotheticals	Conserved hypothetical protein
	Rv1189	sigI	1.65E-05	information pathways	Possible alternative RNA polymerase sigma factor SigI
	Rv0147	-	2.14E-05	intermediary metabolism and respiration	Probable aldehyde dehydrogenase (NAD+) dependent

Rv0963c	-	2.70E-05	conserved hypotheticals intermediary	Conserved hypothetical protein
Rv1988	<i>erm(37)</i>	3.06E-05	metabolism and respiration	Probable 23S rRNA methyltransferase <i>Erm(37)</i>

Drug/A: without adjusting for overlapping resistance; Drug/B: adjusting for overlapping resistance. Loci with grey background were also identified by *phyC*. Locus and gene names in bold indicate known DR loci.

Supplementary Table 11 Operon-based GWAS top hits

Drug/ model	Operon name	P-value	Coding regions and intergenic regions contained in the operon
INH/A	Rv1907c- <i>furA</i>	5.71E-121	Rv1907c, Rv1907c-Rv1908c, <i>katG</i> , <i>katG_promoter</i> , <i>furA</i> , Rv1909c-Rv1910c
	<i>rpoB</i> - <i>rpoB</i>	3.57E-110	Rv0666-Rv0667, <i>rpoB</i>
	<i>embA</i> - <i>embB</i>	7.22E-65	<i>embA_promoter</i> , <i>embA</i> , Rv3794-Rv3795, <i>embB</i>
	Rv1482c-Rv1482c	5.97E-59	Rv1482c
	Rv2037c- <i>pncA</i>	1.14E-58	Rv2037c, Rv2037c-Rv2038c, Rv2038c, Rv2038c-Rv2039c, Rv2039c, Rv2039c-Rv2040c, Rv2040c, Rv2040c-Rv2041c, Rv2041c, Rv2041c-Rv2042c, Rv2042c, Rv2042c-Rv2043c, <i>pncA</i> , Rv2043c-Rv2044c
	<i>fabG1</i> - <i>hemZ</i>	3.09E-49	<i>fabG1</i> , Rv1483-Rv1484, <i>inhA</i> , Rv1484-Rv1485, <i>hemZ</i>
	<i>rpsL</i> - <i>rpsG</i>	3.27E-44	Rv0681-Rv0682, <i>rpsL</i> , Rv0682-Rv0683, <i>rpsG</i>
	Rv1816-Rv1816	4.21E-36	Rv1815-Rv1816, Rv1816
INH/B	Rv1907c- <i>furA</i>	1.09E-43	Rv1907c, Rv1907c-Rv1908c, <i>katG</i> , <i>katG_promoter</i> , <i>furA</i> , Rv1909c-Rv1910c
	Rv1482c-Rv1482c	1.16E-17	Rv1482c
	<i>fabG1</i> - <i>hemZ</i>	2.61E-16	<i>fabG1</i> , Rv1483-Rv1484, <i>inhA</i> , Rv1484-Rv1485, <i>hemZ</i>
	<i>pgmA</i> - <i>pgmA</i>	2.50E-12	<i>pgmA</i>
	<i>rpsS</i> - <i>rpsQ</i>	1.23E-11	Rv0704-Rv0705, <i>rpsS</i> , Rv0705-Rv0706, <i>rplV</i> , Rv0706-Rv0707, <i>rpsC</i> , Rv0707-Rv0708, <i>rplP</i> , Rv0708-Rv0709, <i>rpmC</i> , Rv0709-Rv0710, <i>rpsQ</i>
	<i>end</i> - <i>lpqP</i>	2.38E-11	<i>end</i> , Rv0670-Rv0671, <i>lpqP</i>
RMP/A	<i>pckA</i> - <i>pckA</i>	1.36E-25	Rv0210-Rv0211, <i>pckA</i>
	<i>ctpl</i> - <i>ctpl</i>	7.73E-21	<i>ctpl</i> , Rv0107c-Rv0108c
	Rv0147-Rv0147	8.86E-20	Rv0146-Rv0147, Rv0147
	<i>yrbE1A</i> -Rv0178	4.80E-19	Rv0166-Rv0167, <i>yrbE1A</i> , Rv0167-Rv0168, <i>yrbE1B</i> , Rv0168-Rv0169, <i>mce1A</i> , Rv0169-Rv0170, <i>mce1B</i> , Rv0170-Rv0171, <i>mce1C</i> , Rv0171-Rv0172, <i>mce1D</i> , Rv0172-Rv0173, <i>lprK</i> , Rv0173-Rv0174, <i>mce1F</i> , Rv0174-Rv0175, Rv0175, Rv0175-Rv0176, Rv0176, Rv0176-Rv0177, Rv0177, Rv0177-Rv0178, Rv0178
	Rv0143c-Rv0143c	1.55E-18	Rv0143c
	<i>rpsR</i> - <i>dnaB</i>	2.34E-13	Rv0054-Rv0055, <i>rpsR1</i> , Rv0055-Rv0056, <i>rplI</i> , Rv0056-Rv0057, Rv0057, Rv0057-Rv0058, <i>dnaB</i>
	Rv0023-Rv0025	5.24E-11	Rv0023, Rv0023-Rv0024, Rv0024, Rv0024-Rv0025, Rv0025
	<i>mmpL3</i> - <i>mmpL3</i>	2.31E-09	<i>mmpL3</i> , Rv0206c-Rv0207c

	Rv0205-Rv0205	3.08E-09	Rv0205
	PE_PGRS2- PE_PGRS2	3.43E-09	Rv0123-Rv0124, PE_PGRS2
RMP/B	rpoB-rpoB	1.68E-05	Rv0666-Rv0667, rpoB
	embA-embB	2.12E-24	embA_promoter, embA, Rv3794-Rv3795, embB
	rpoB-rpoB	4.31E-15	Rv0666-Rv0667, rpoB
	Rv2037c-pncA	5.57E-13	Rv2037c, Rv2037c-Rv2038c, Rv2038c, Rv2038c- Rv2039c, Rv2039c, Rv2039c-Rv2040c, Rv2040c, Rv2040c-Rv2041c, Rv2041c, Rv2041c-Rv2042c, Rv2042c, Rv2042c-Rv2043c, pncA, Rv2043c-Rv2044c
EMB/A	Rv0818-Rv0819	1.37E-08	Rv0818, Rv0818-Rv0819, mshD
	Rv1907c-furA	1.33E-07	Rv1907c, Rv1907c-Rv1908c, katG, katG_promoter, furA, Rv1909c-Rv1910c
	cspB-cspB	1.53E-06	cspB
	moaD2-Rv0870c	2.45E-06	moaD2, Rv0868c-Rv0869c, moaA2, Rv0869c-Rv0870c, Rv0870c
	rpsL-rpsG	3.29E-06	Rv0681-Rv0682, rpsL, Rv0682-Rv0683, rpsG
	Rv0726c-Rv0726c	6.11E-06	Rv0726c, Rv0726c-Rv0727c
	embA-embB	1.87E-05	embA_promoter, embA, Rv3794-Rv3795, embB
	ugpC-ugpA	0.005977	ugpC, Rv2832c-Rv2833c, ugpB, Rv2833c-Rv2834c, ugpE, Rv2834c-Rv2835c, ugpA, Rv2835c-Rv2836c
	gpsA-gpsA	0.008341	gpdA1, Rv0564c-Rv0565c, gpdA2
EMB/B	cyp139-Rv1668c	0.008609	cyp139, Rv1666c-Rv1667c, Rv1667c, Rv1667c-Rv1668c, Rv1668c
	atpA-Rv1312	0.009098	Rv1307-Rv1308, atpA, Rv1308-Rv1309, atpG, Rv1309- Rv1310, atpD, Rv1310-Rv1311, atpC, Rv1311-Rv1312, Rv1312
	PE15-PPE20	0.009489	Rv1385-Rv1386, PE15, Rv1386-Rv1387, PPE20
	rpoB-rpoB	3.89E-35	Rv0666-Rv0667, rpoB
	fabG1-hemZ	1.06E-19	fabG1, Rv1483-Rv1484, inhA, Rv1484-Rv1485, hemZ
	pks3-pks3	2.28E-18	pks3
	Rv1482c-Rv1482c	2.49E-18	Rv1482c
	rpsL-rpsG	2.29E-14	Rv0681-Rv0682, rpsL, Rv0682-Rv0683, rpsG
PZA/A	Rv1179c-Rv1179c	4.94E-14	Rv1179c
	PE_PGRS28- PE_PGRS28	1.56E-13	PE_PGRS28
	Rv1148c-Rv1148c	2.95E-12	Rv1148c
	rpoC-rpoC	5.30E-11	Rv0667-Rv0668, rpoC
	kdpF-kdpC	6.20E-11	kdpF, Rv1028A-Rv1029, kdpA, Rv1029-Rv1030, kdpB, Rv1030-Rv1031, kdpC
	Rv2037c-pncA	0.000389	Rv2037c, Rv2037c-Rv2038c, Rv2038c, Rv2038c- Rv2039c, Rv2039c, Rv2039c-Rv2040c, Rv2040c, Rv2040c-Rv2041c, Rv2041c, Rv2041c-Rv2042c, Rv2042c, Rv2042c-Rv2043c, pncA, Rv2043c-Rv2044c
PZA/B	PPE55-PPE55	0.000495	PPE55
	Rv2295-Rv2295	0.001133	pncA_promoter, Rv2295
	PE_PGRS22- PE_PGRS22	0.001293	Rv1090-Rv1091, PE_PGRS22
	PE_PGRS7-	0.003927	PE_PGRS7

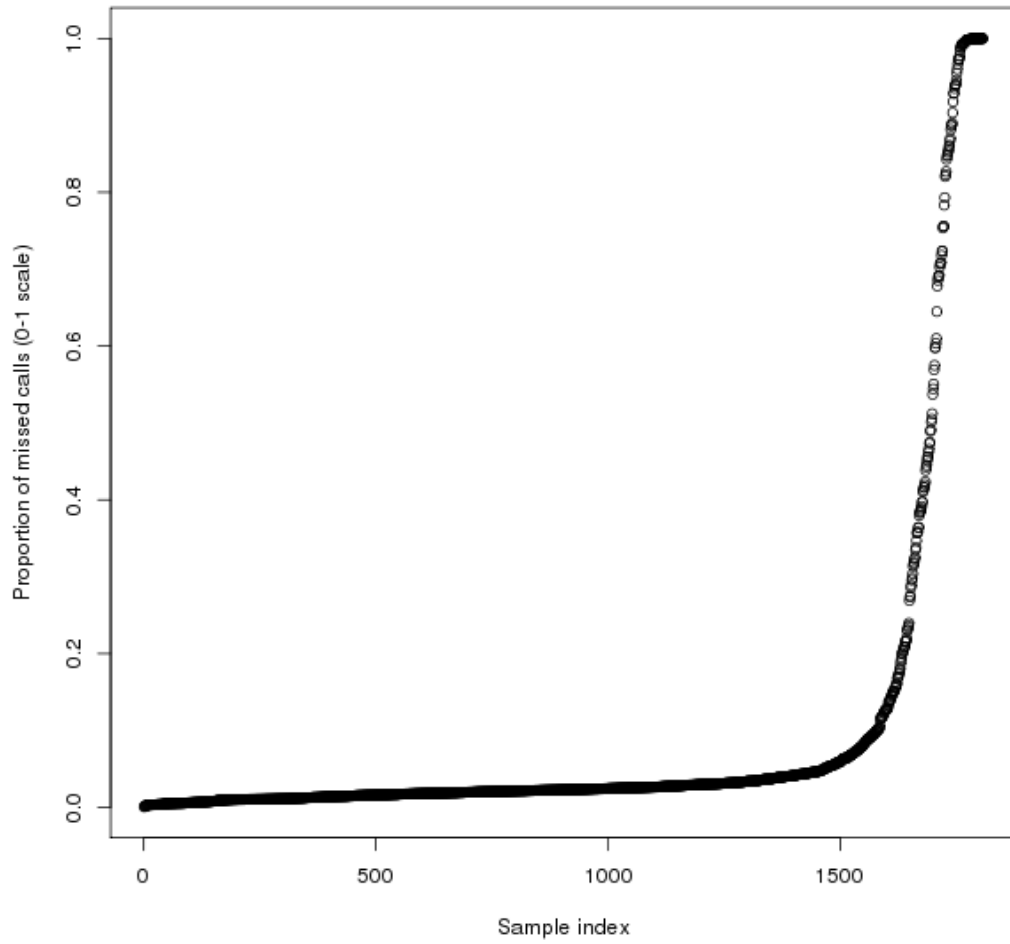
	PE_PGRS7		
	PE_PGRS43- PE_PGRS43	0.004543	PE_PGRS43
	Rv0104-Rv0104	0.00739	Rv0104
	PE_PGRS3- PE_PGRS3	0.00858	PE_PGRS3, Rv0278c-Rv0279c
	rpsL-rpsG	4.78E-49	Rv0681-Rv0682, rpsL, Rv0682-Rv0683, rpsG
	rpoB-rpoB	1.50E-32	Rv0666-Rv0667, rpoB
	fabG1-hemZ	4.69E-31	fabG1, Rv1483-Rv1484, inhA, Rv1484-Rv1485, hemZ
	Rv1482c-Rv1482c	5.30E-25	Rv1482c
	Rv0143c-Rv0143c	4.12E-23	Rv0143c
STR/A	pckA-pckA	1.50E-19	Rv0210-Rv0211, pckA
	Rv1816-Rv1816	1.60E-18	Rv1815-Rv1816, Rv1816
	Rv1998c-Rv1998c	1.61E-18	Rv1998c, Rv1998c-Rv1999c
	Rv1148c-Rv1148c	1.74E-18	Rv1148c
	Rv0947c-Rv0948c	1.75E-18	Rv0947c, Rv0947c-Rv0948c, Rv0948c
	rpsL-rpsG	1.70E-06	Rv0681-Rv0682, rpsL, Rv0682-Rv0683, rpsG
	Rv2595-Rv2596	1.83E-05	vapB40, Rv2595-Rv2596, vapC40
	Rv0147-Rv0147	2.50E-05	Rv0146-Rv0147, Rv0147
	Rv0302-Rv0303	2.67E-05	Rv0301-Rv0302, Rv0302, Rv0302-Rv0303, Rv0303
STR/B	pknD-pstS2	2.68E-05	pknD, Rv0931c-Rv0932c, pstS2
	Rv0963c-Rv0963c	2.68E-05	Rv0963c, Rv0963c-Rv0964c
	Rv2629-Rv2630	2.83E-05	Rv2628-Rv2629, Rv2629, Rv2629-Rv2630, Rv2630
	Rv3728-Rv3728	2.86E-05	Rv3727-Rv3728, Rv3728
	pgsA3-pgsA3	3.15E-05	pgsA3
	rpoB-rpoB	3.54E-32	Rv0666-Rv0667, rpoB
	aspC-Rv0338c	2.59E-25	aspC, Rv0337c-Rv0338c, Rv0338c, Rv0338c-Rv0339c
	Rv0147-Rv0147	1.60E-24	Rv0146-Rv0147, Rv0147
	pckA-pckA	1.60E-24	Rv0210-Rv0211, pckA
	Rv0143c-Rv0143c	3.99E-24	Rv0143c
ETH/A	Rv0276-Rv0276	6.20E-24	Rv0276
	pknD-pstS2	3.66E-22	pknD, Rv0931c-Rv0932c, pstS2
	rpsL-rpsG	8.01E-22	Rv0681-Rv0682, rpsL, Rv0682-Rv0683, rpsG
	phoR-phoR	9.26E-21	Rv0757-Rv0758, phoR
	Rv0302-Rv0303	9.58E-21	Rv0301-Rv0302, Rv0302, Rv0302-Rv0303, Rv0303
	fabG1-hemZ	0.000518	fabG1, Rv1483-Rv1484, inhA, Rv1484-Rv1485, hemZ
	Rv1816-Rv1816	0.000674	Rv1815-Rv1816, Rv1816
	Rv0194-Rv0194	0.000842	Rv0194
	Rv1482c-Rv1482c	0.000979	Rv1482c
ETH/B	metH-metH	0.005044	metH
	fadD11.1-plsB1	0.010402	fadD11.1, Rv1549-Rv1550, fadD11, Rv1550-Rv1551, plsB1
	Rv2075c-Rv2075c	0.011742	Rv2075c, Rv2075c-Rv2076c
	Rv1873-Rv1873	0.01181	Rv1873
	Rv3915-Rv3915	0.013046	Rv3914-Rv3915, Rv3915

	Rv3654c-Rv3659c	0.014992	Rv3654c, Rv3654c-Rv3655c, Rv3655c, Rv3655c-Rv3656c, Rv3656c, Rv3656c-Rv3657c, Rv3657c, Rv3657c-Rv3658c, Rv3658c, Rv3658c-Rv3659c, Rv3659c, Rv3659c-Rv3660c
	embA-embB	9.13E-15	embA_promoter, embA, Rv3794-Rv3795, embB
	gyrB-gyrA	7.95E-11	Rv0004-Rv0005, gyrB, Rv0005-Rv0006, gyrA
	Rv1482c-Rv1482c	4.94E-09	Rv1482c
	fabG1-hemZ	9.67E-09	fabG1, Rv1483-Rv1484, inhA, Rv1484-Rv1485, hemZ
OFX/A	ccrB-Rv3071	1.65E-07	Rv3069, Rv3069-Rv3070, Rv3070, Rv3070-Rv3071, Rv3071
	eis-eis	1.56E-06	eis, eis_promoter
	gcp-alr	8.81E-06	gcp, Rv3419c-Rv3420c, rimI, Rv3420c-Rv3421c, Rv3421c, Rv3421c-Rv3422c, Rv3422c, Rv3422c-Rv3423c, alr, Rv3423c-Rv3424c
	gyrB-gyrA	0.013219	Rv0004-Rv0005, gyrB, Rv0005-Rv0006, gyrA
	Rv1635c-Rv1635c	0.017087	Rv1635c
	mmpL3-mmpL3	0.026943	mmpL3, Rv0206c-Rv0207c
	Rv3224-Rv3224B	0.034723	Rv3224, Rv3224-Rv3224A, Rv3224A, Rv3224A-Rv3224B, Rv3224B
	mprA-mprB	0.039519	mprA, Rv0981-Rv0982, mprB
OFX/B	Rv3090-Rv3091	0.042676	Rv3089-Rv3090, Rv3090, Rv3090-Rv3091, Rv3091
	PE_PGRS18-PE_PGRS18	0.04297	PE_PGRS18
	pkS7-pkS9	0.051017	Rv1660-Rv1661, pkS7, Rv1661-Rv1662, pkS8, Rv1662-Rv1663, pkS17, Rv1663-Rv1664, pkS9
	PPE35-PPE35	0.051095	PPE35, Rv1918c-Rv1919c
	PPE1-nrp	0.052266	PPE1, Rv0096-Rv0097, Rv0097, Rv0097-Rv0098, fcoT, Rv0098-Rv0099, fadD10, Rv0099-Rv0100, Rv0100, Rv0100-Rv0101, nrp
	atpA-Rv1312	1.56E-11	Rv1307-Rv1308, atpA, Rv1308-Rv1309, atpG, Rv1309-Rv1310, atpD, Rv1310-Rv1311, atpC, Rv1311-Rv1312, Rv1312
	uvrC-whiA	6.85E-11	Rv1419-Rv1420, uvrC, Rv1420-Rv1421, Rv1421, Rv1421-Rv1422, Rv1422, Rv1422-Rv1423, whiA
	bioB-Rv1591	1.42E-10	bioB, Rv1589-Rv1590, Rv1590, Rv1590-Rv1591, Rv1591
	Rv1697-Rv1698	1.72E-10	Rv1696-Rv1697, Rv1697, Rv1697-Rv1698, mctB
AMK/A	Rv1137c-Rv1139c	2.74E-10	Rv1137c, Rv1137c-Rv1138c, Rv1138c, Rv1138c-Rv1139c, Rv1139c
	Rv1959c-Rv1960c	5.76E-10	parE1, Rv1959c-Rv1960c, parD1
	Rv1140-Rv1140	1.21E-09	Rv1140
	Rv0613c-Rv0613c	1.39E-09	Rv0613c
	Rv1065-Rv1066	1.07E-08	Rv1065, Rv1065-Rv1066, Rv1066
	Rv1004c-Rv1004c	2.17E-08	Rv1004c, Rv1004c-Rv1005c
	embA-embB	8.46E-06	embA_promoter, embA, Rv3794-Rv3795, embB
	Rv2181-Rv2181	2.44E-05	Rv2181
AMK/B	ugpC-ugpA	2.59E-05	ugpC, Rv2832c-Rv2833c, ugpB, Rv2833c-Rv2834c, ugpE, Rv2834c-Rv2835c, ugpA, Rv2835c-Rv2836c
	deoD-pmmB	2.69E-05	deoD, Rv3307-Rv3308, pmmB
	Rv3182-Rv3183	4.97E-05	Rv3182, Rv3182-Rv3183, Rv3183

	lppR-lepA	5.60E-05	lppR, Rv2403c-Rv2404c, lepA
	atpA-Rv1312	6.04E-05	Rv1307-Rv1308, atpA, Rv1308-Rv1309, atpG, Rv1309-Rv1310, atpD, Rv1310-Rv1311, atpC, Rv1311-Rv1312, Rv1312
	Rv1482c-Rv1482c	9.84E-10	Rv1482c
	fabG1-hemZ	2.78E-09	fabG1, Rv1483-Rv1484, inhA, Rv1484-Rv1485, hemZ
	Rv0573c-Rv0574c	3.21E-08	pncB2, Rv0573c-Rv0574c, Rv0574c, Rv0574c-Rv0575c
	Rv1140-Rv1140	7.31E-08	Rv1140
	trcS-trcR	9.20E-08	trcS, Rv1032c-Rv1033c, trcR, Rv1033c-Rv1034c
CAP/A	Rv1697-Rv1698	1.07E-07	Rv1696-Rv1697, Rv1697, Rv1697-Rv1698, mctB
	Rv0148-Rv0149	1.17E-07	Rv0147-Rv0148, Rv0148, Rv0148-Rv0149, Rv0149
	atpA-Rv1312	2.02E-07	Rv1307-Rv1308, atpA, Rv1308-Rv1309, atpG, Rv1309-Rv1310, atpD, Rv1310-Rv1311, atpC, Rv1311-Rv1312, Rv1312
	Rv0613c-Rv0613c	3.43E-07	Rv0613c
	Rv0818-Rv0819	8.06E-07	Rv0818, Rv0818-Rv0819, mshD
	embA-embB	6.43E-08	embA_promoter, embA, Rv3794-Rv3795, embB
	Rv1482c-Rv1482c	7.46E-06	Rv1482c
	fabG1-hemZ	1.11E-05	fabG1, Rv1483-Rv1484, inhA, Rv1484-Rv1485, hemZ
CAP/B	Rv3182-Rv3183	9.90E-05	Rv3182, Rv3182-Rv3183, Rv3183
	mrp-Rv1230c	0.000125	mrp, Rv1229c-Rv1230c, Rv1230c, Rv1230c-Rv1231c
	Rv0148-Rv0149	0.000214	Rv0147-Rv0148, Rv0148, Rv0148-Rv0149, Rv0149
	embA-embB	3.23E-13	embA_promoter, embA, Rv3794-Rv3795, embB
KAN/A	eis-eis	1.07E-10	eis, eis_promoter
	Rv2075c-Rv2075c	8.66E-07	Rv2075c, Rv2075c-Rv2076c
	Rv2313c-Rv2315c	6.42E-06	Rv2313c, Rv2313c-Rv2314c, Rv2314c, Rv2314c-Rv2315c, Rv2315c
	Rv3067-Rv3067	6.56E-06	Rv3066-Rv3067, Rv3067
KAN/B	eis-eis	1.01E-05	eis, eis_promoter
	pstB-pstS1	1.53E-05	pstB, Rv0933-Rv0934, pstS1
	Rv1988-Rv1988	1.91E-05	Rv1987-Rv1988, erm(37)
	Rv0963c-Rv0963c	2.50E-05	Rv0963c, Rv0963c-Rv0964c

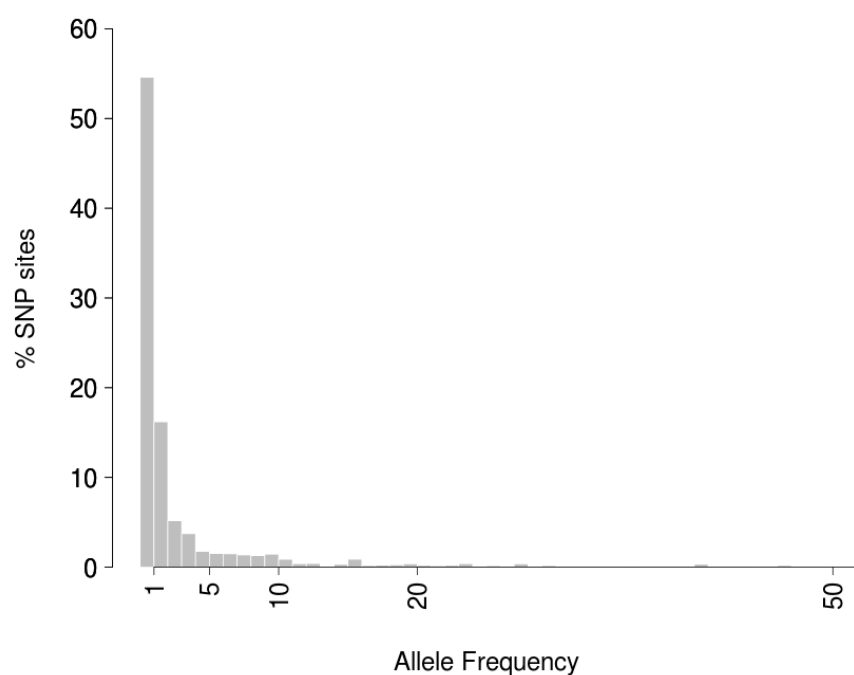
Drug/A: without adjusting for overlapping resistance; Drug/B: adjusting for overlapping resistance. Operon annotation was extracted from TBDB (Reddy *et al.* 2009). The forth column includes the name of coding and intergenic regions operons are composed of.

Supplementary Figure 1 Proportion of missed call across all samples



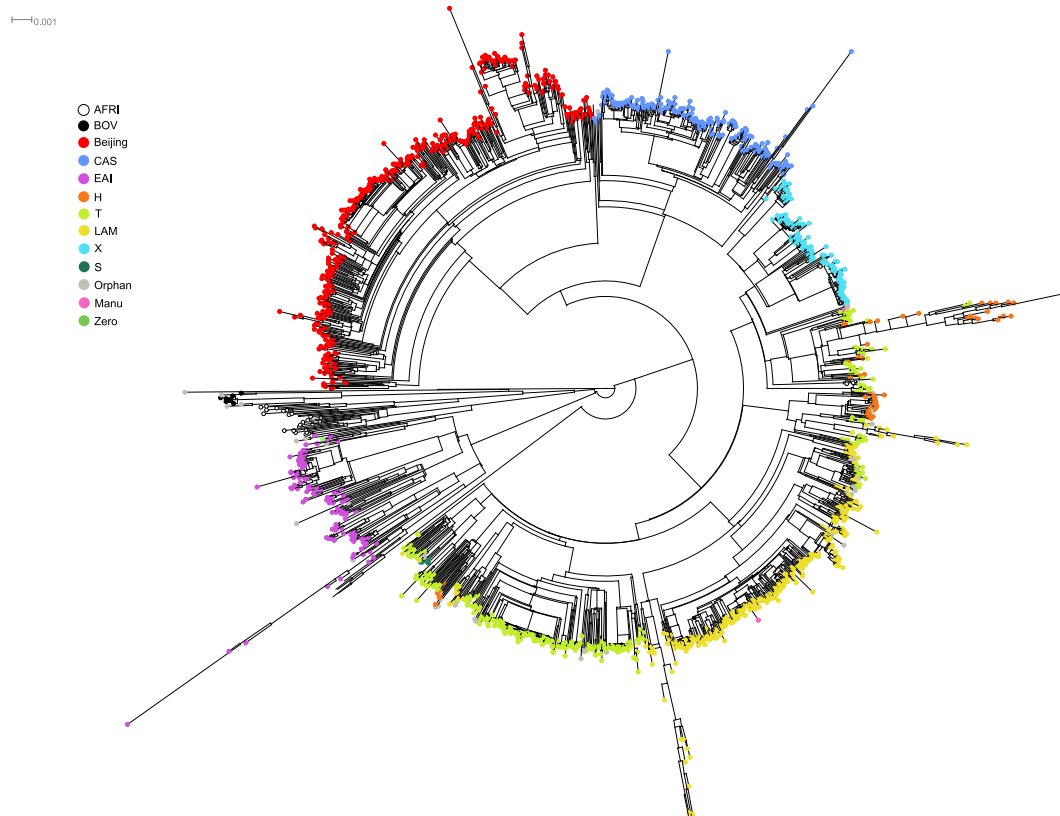
The proportion of missed calls (i.e. SNP alleles that could not be called due to low coverage) is ordered and plotted for all samples. An inflexion point is observed at 0.15 (15%) and used to filter out bad quality samples.

Supplementary Figure 2 SNP Allele Frequency Spectrum



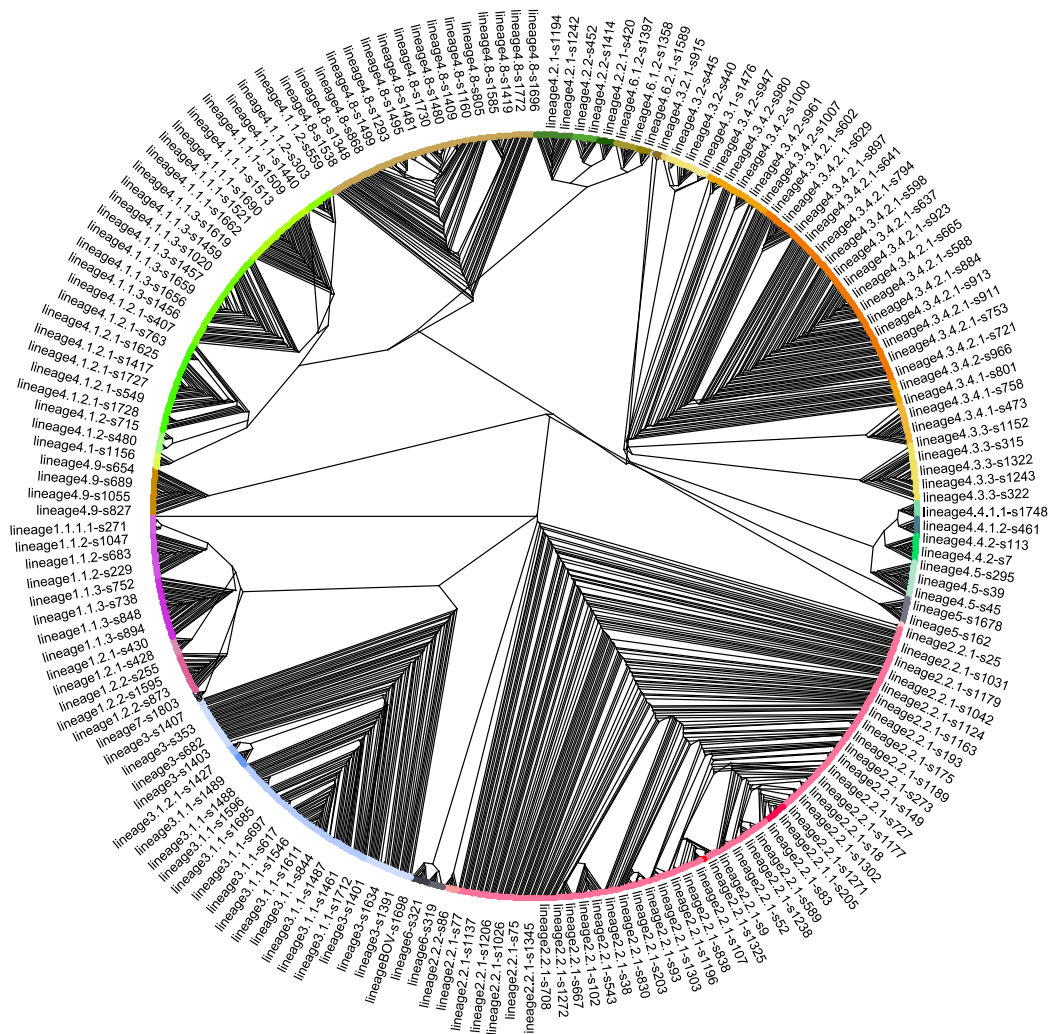
The allele frequency (x-axis) refers to the number of samples across the whole data set sharing a particular SNP. The first bar represents the percentage of SNPs (y-axis) (out of 91,648) present in only one sample and absent in the rest, i.e. private SNPs. The second bar presents the percentage of SNPs harboured by two different samples, and so on.

Supplementary Figure 3 Global phylogeny of 1,601 MTBC isolates colour, coded by spoligotype



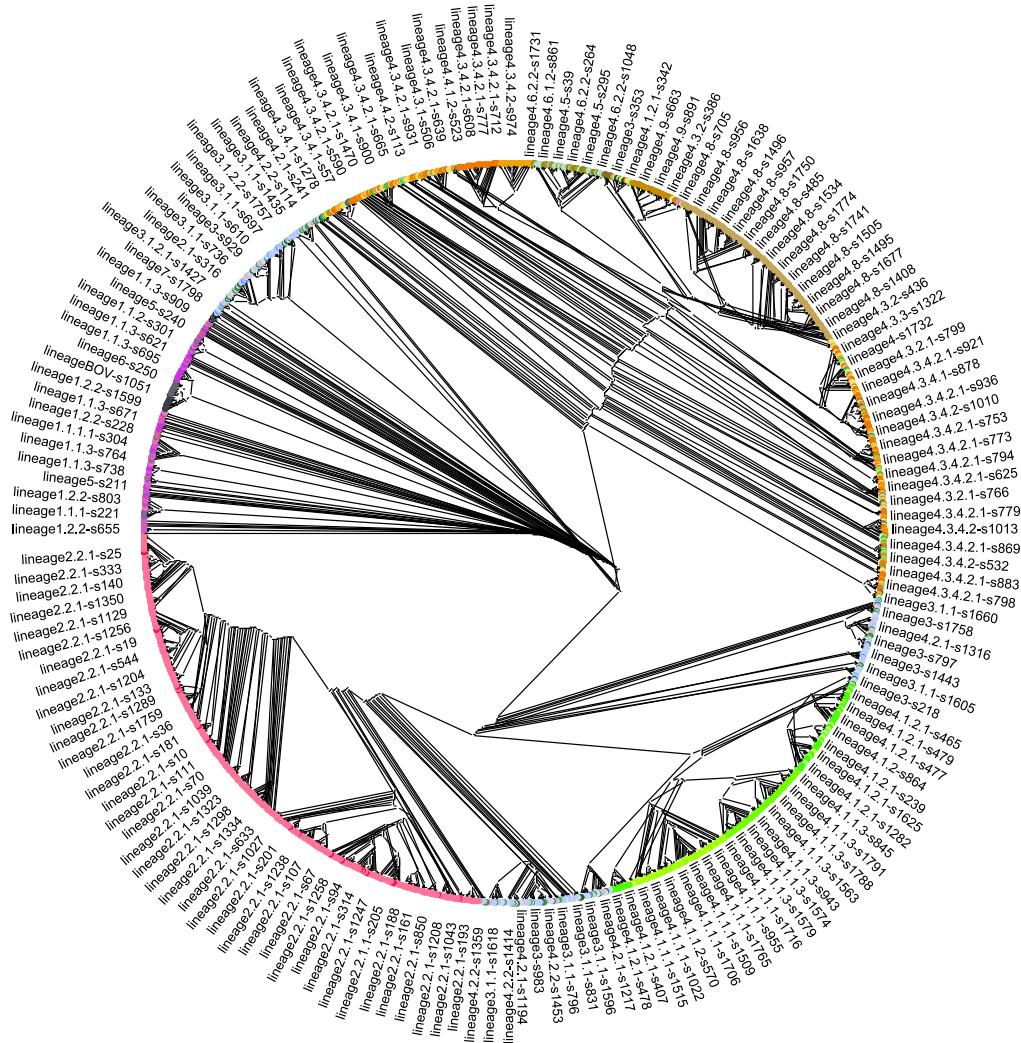
Spoligotypes families generally cluster within specific SNP, defined clades, as it is the case for Beijing, CAS, EAI, AFRI_1, AFRI_2, BOV, X and S. However, there is evidence of homoplasy particularly in lineage 4, among T, H and LAM spoligotypes.

Supplementary Figure 4 Global phylogeny constructed using the 62 SNP typing system



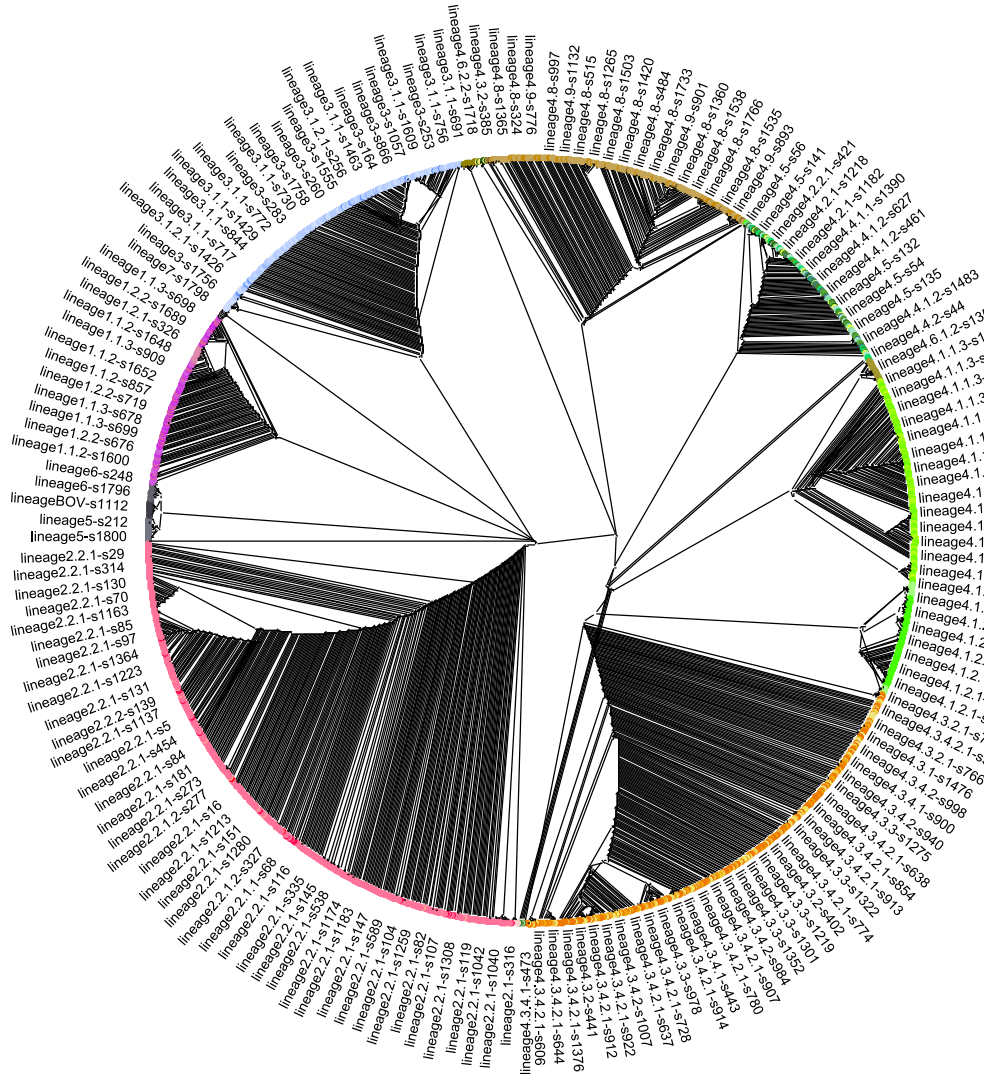
Global phylogeny (WGS data set 2; n=1,601 samples) constructed using only the proposed minimum set of 62 SNPs separates all 1,601 samples into their corresponding lineage and sub-lineage.

Supplementary Figure 5 Global phylogeny constructed using the 45 SNP typing system proposed by Filliol *et al.* 2006



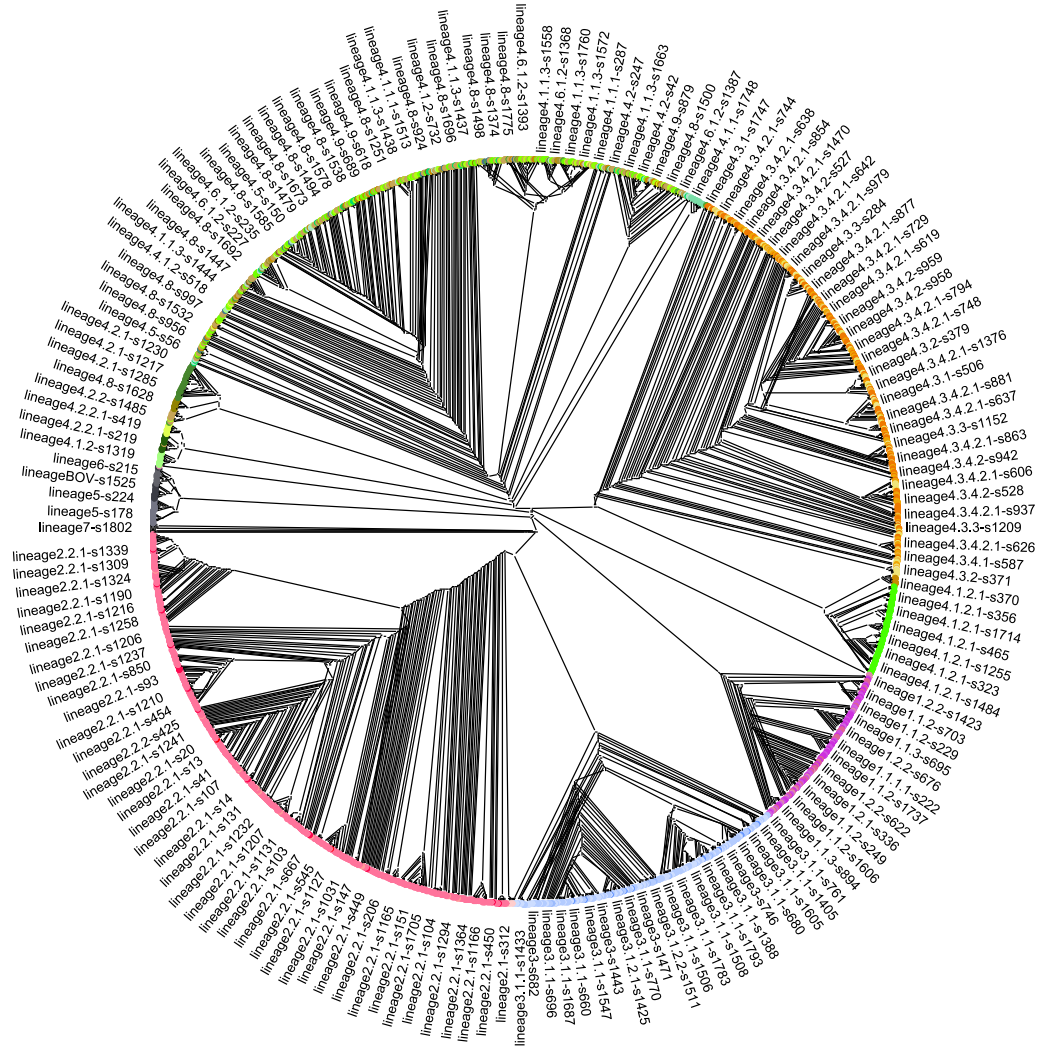
Global phylogeny (WGS data set 2; n=1,601 samples) constructed using the set of 45 SNPs proposed by (Filliol *et al.* 2006) yielded an incompatible classification compared to that obtained by Comas93 (Supplementary Figure 6) and Homolka71 (Supplementary Figure 7). The MTBC lineages could not be unambiguously separated (lineages 1, 5, 6 and *M. bovis*) or were spread across multiple clades (lineages 2, 3 and 4).

Supplementary Figure 6 Global phylogeny constructed using the 93 SNP typing system proposed by Comas *et al.* 2009



A global phylogeny (WGS data set 2; n=1,601 samples) constructed using the 93 lineage-specific SNPs proposed by (Comas *et al.* 2009) shows all 6 MTBC main lineages and *M. bovis* unambiguously separated in different clades. Samples from sub-lineage 2.1 (non-Beijing), 2.2 (Beijing), 4.3 (LAM), 4.1.1 (X-family), 4.1.2.1 (Haarlem), 4.6.2.2 (Cameroon) and 4.6.1 (Uganda) were all constrained to specific clades. However, the markers lacked resolution at a sub-lineage level, with the majority being unresolvable.

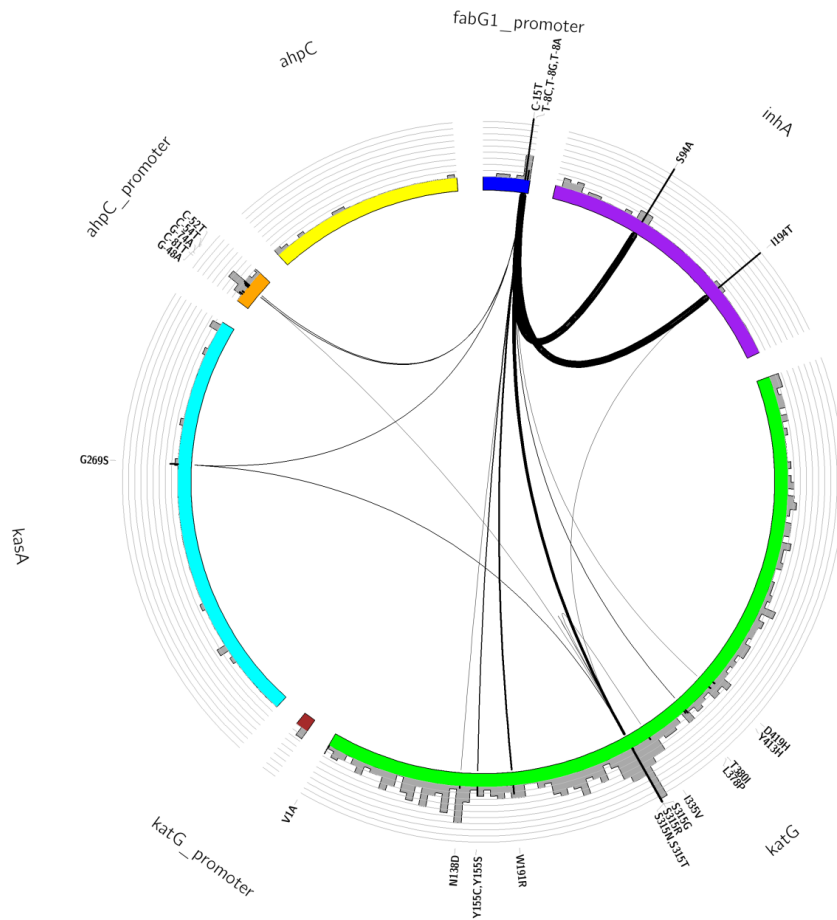
Supplementary Figure 7 Global phylogeny constructed using the 71 SNP typing system proposed by Homolka *et al.* 2012



Global phylogeny constructed (WGS data set 2; n=1,601 samples) using the 71 phylogenetically informative SNPs proposed by (Homolka 2012). The phylogeny is largely congruent with that built using Comas93 SNP set, all main seven MTBC lineages are clearly separated. Samples belonging to sub-lineages 4.2.1 (Ural), 4.2.2.1 (TUR), 4.3 (LAM), 4.4.1.1 (S-type), 4.1.2.1 (Haarlem) and 4.6.2.2 (Cameroon) are all restricted to specific clades. However other sub-lineages (particularly from lineage 4) were not congruent with the RD system with some samples with the same RD (e.g. RD115 for 4.3.3 sub-lineages) spread across different clades.

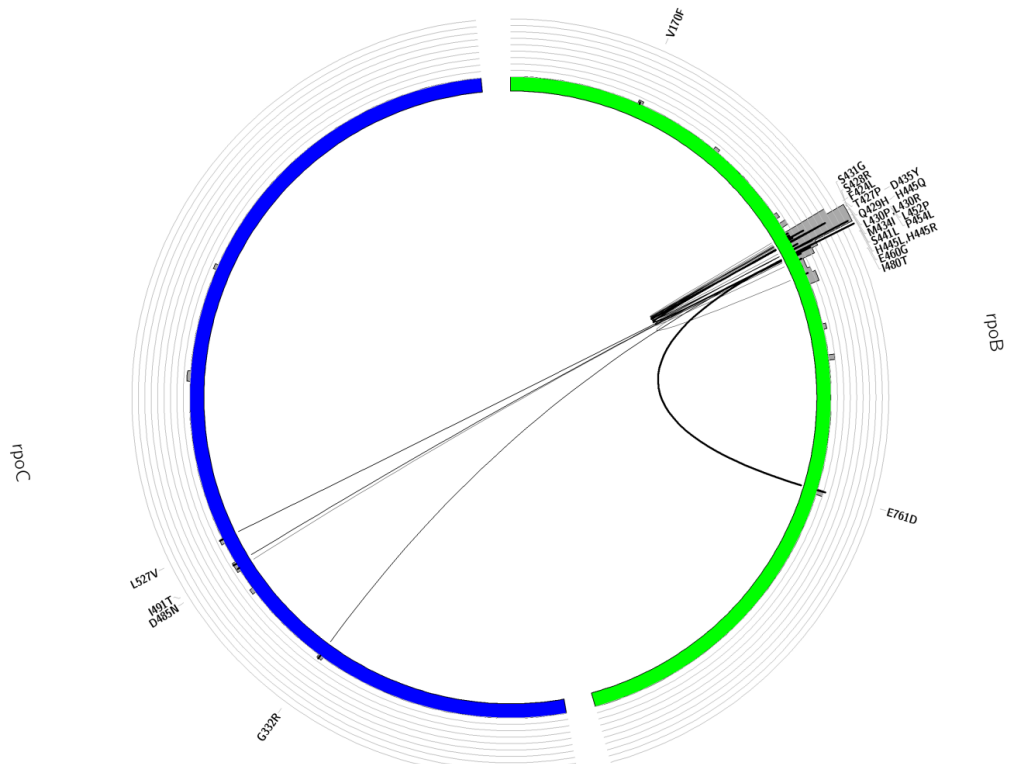
Supplementary Figure 8 Loci involved in drug resistance

(A) Loci involved in isoniazid resistance and mutations observed in isoniazid resistant isolates



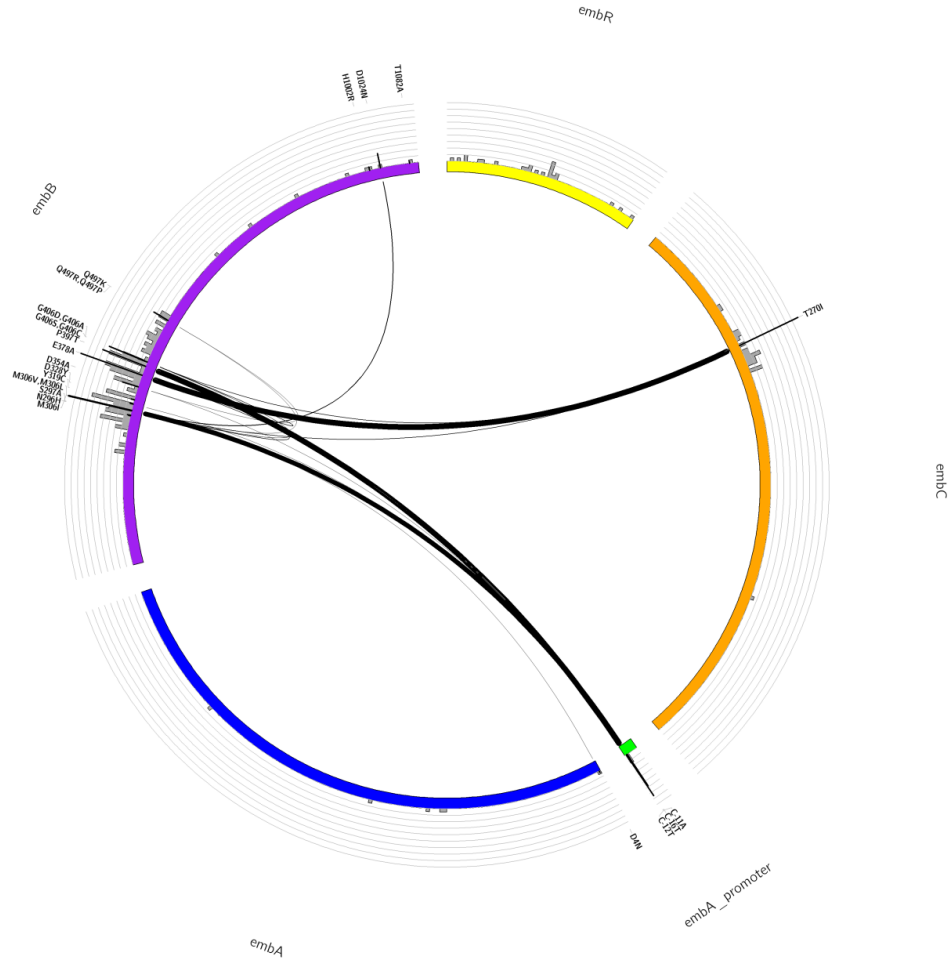
Summary of DR-associated genes and mutations in the curated library for INH (296 variable sites, 350 SNPs and 25 indels, in 4 genes and 3 promoters) and mutations observed in phenotypically INH resistant samples. Colour-coded bars in the Circos plot represent genes described to be involved in DR. On top of each of these bars a grey histogram shows the mutation density derived from the curated list of DR-associated mutations. These grey areas highlight the presence of DR-associated regions in candidate genes. Vertical black lines indicate the frequency of mutations observed in phenotypically resistance isolates. Internal black lines show co-occurring mutations both within and between genes. The thickness of these lines is proportional to the frequency of the mutations appearing together.

(B) Loci involved in rifampicin resistance and mutations observed in rifampicin resistant isolates



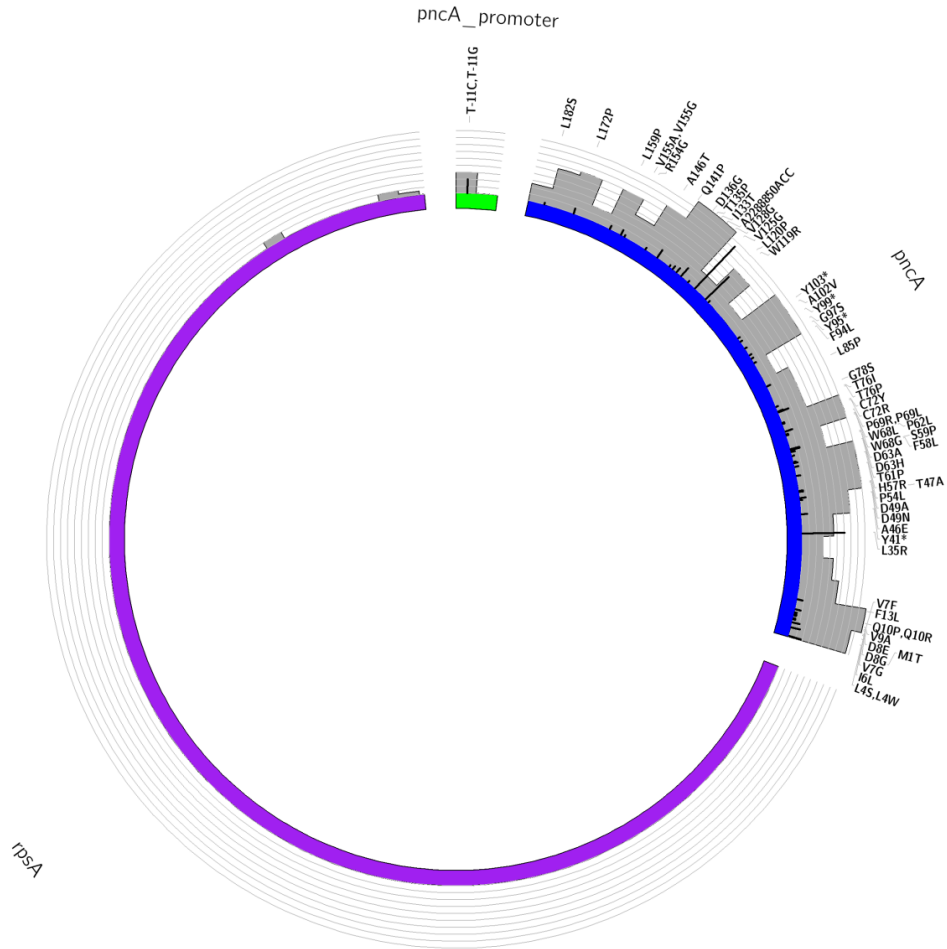
Summary of DR-associated genes and mutations in the curated library for RMP (97 variable sites, 143 SNPs and 19 indels, in 2 genes) and mutations observed in phenotypically RMP resistant samples. Colour-coded bars in the Circos plot represent genes described to be involved in DR. On top of each of these bars a grey histogram shows the mutation density derived from the curated list of DR-associated mutations. These grey areas highlight the presence of DR-associated regions in candidate genes. Vertical black lines indicate the frequency of mutations observed in phenotypically resistance isolates. Internal black lines show co-occurring mutations both within and between genes. The thickness of these lines is proportional to the frequency of the mutations appearing together.

(C) Loci involved in ethambutol resistance and mutations observed in ethambutol resistant isolates



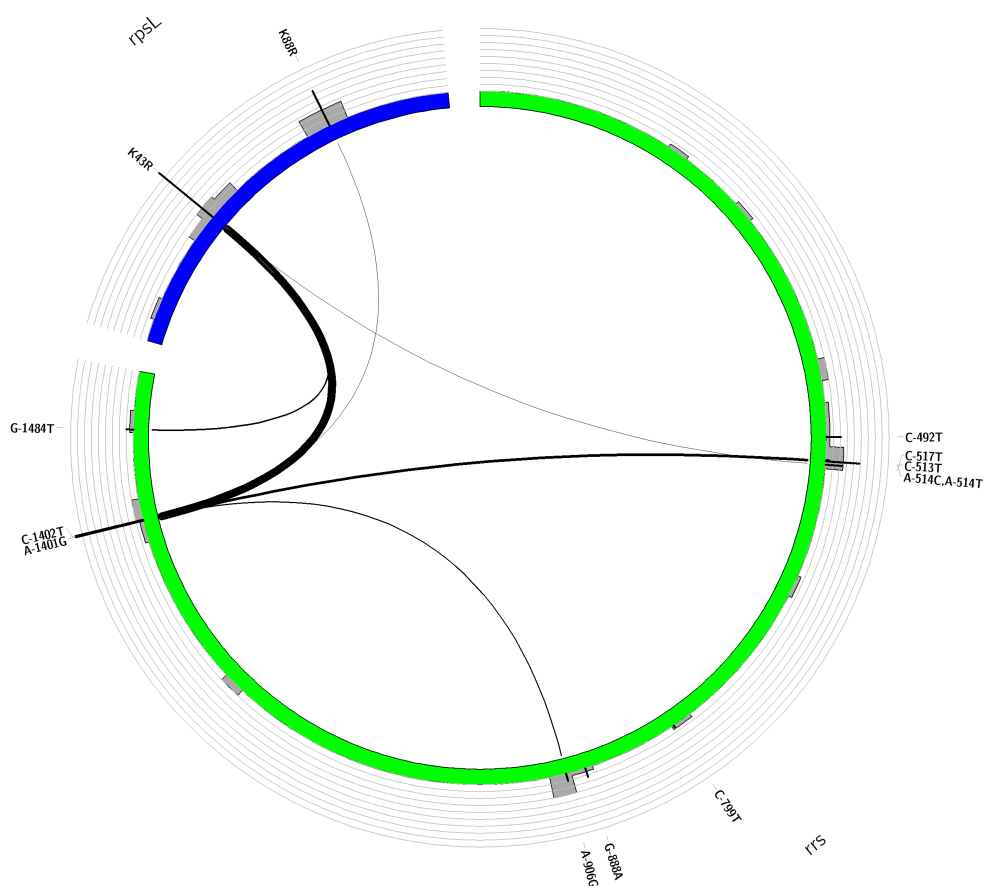
Summary of DR-associated genes and mutations in the curated library for EMB (180 variable sites, 213 SNPs and 1 indel, in 4 genes and 1 promoter) and mutations observed in phenotypically EMB resistant samples. Colour-coded bars in the Circos plot represent genes described to be involved in DR. On top of each of these bars a grey histogram shows the mutation density derived from the curated list of DR-associated mutations. These grey areas highlight the presence of DR-associated regions in candidate genes. Vertical black lines indicate the frequency of mutations observed in phenotypically resistance isolates. Internal black lines show co-occurring mutations both within and between genes. The thickness of these lines is proportional to the frequency of the mutations appearing together.

(D) Loci involved in pyrazinamide resistance and mutations observed in pyrazinamide resistant isolates



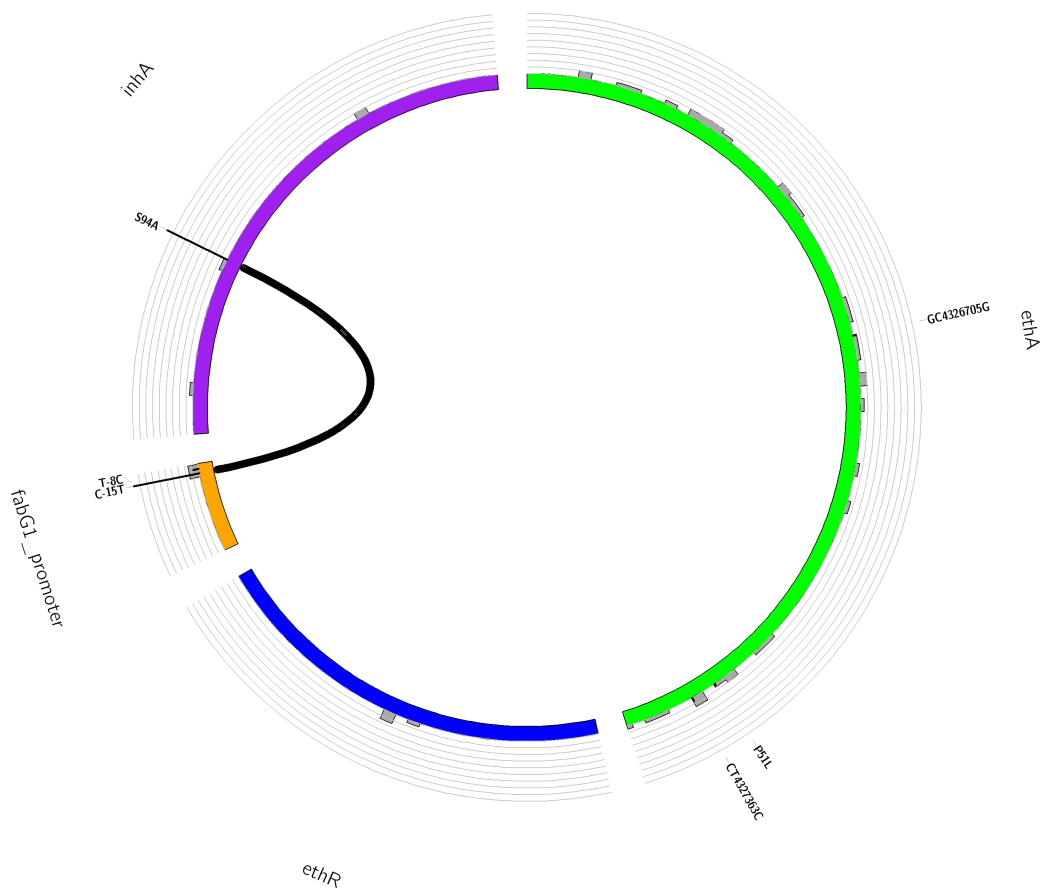
Summary of DR-associated genes and mutations in the curated library for PZA (225 variable sites, 280 SNPs and 64 indels, in 2 genes and 1 promoter) and mutations observed in phenotypically PZA resistant samples. Colour-coded bars in the Circos plot represent genes described to be involved in DR. On top of each of these bars a grey histogram shows the mutation density derived from the curated list of DR-associated mutations. These grey areas highlight the presence of DR-associated regions in candidate genes. Vertical black lines indicate the frequency of mutations observed in phenotypically resistance isolates. Internal black lines show co-occurring mutations both within and between genes. The thickness of these lines is proportional to the frequency of the mutations appearing together.

(E) Loci involved in streptomycin resistance and mutations observed in streptomycin resistant isolates



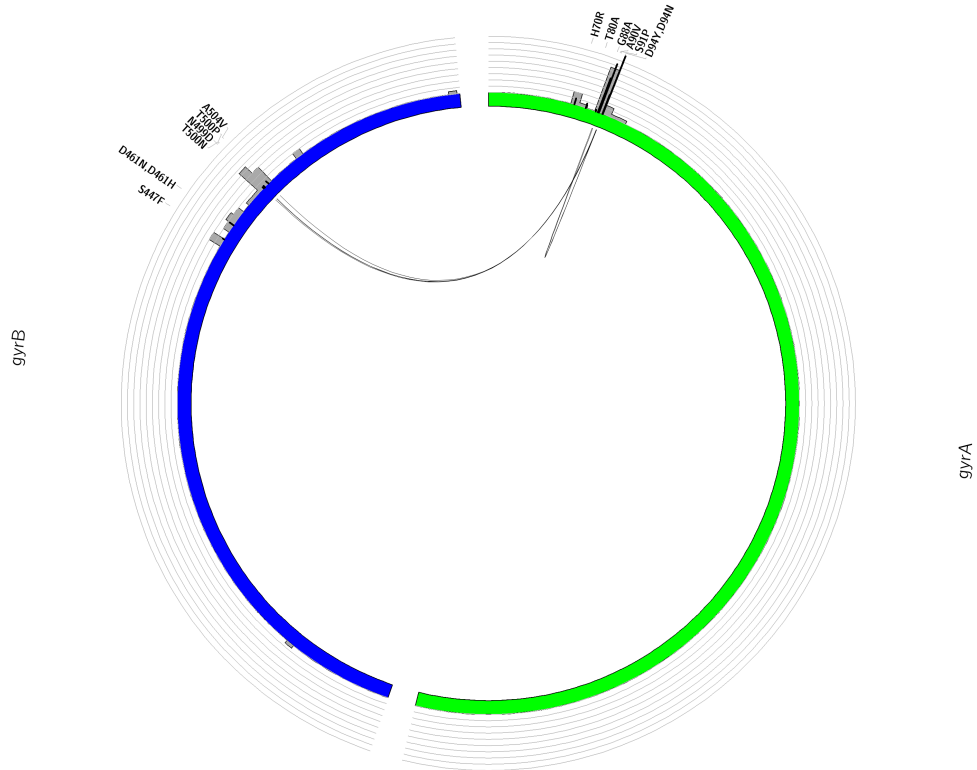
Summary of DR-associated genes and mutations in the curated library for STR (35 variable sites, 44 SNPs, in 2 genes) and mutations observed in phenotypically STR resistant samples. Colour-coded bars in the Circos plot represent genes described to be involved in DR. On top of each of these bars a grey histogram shows the mutation density derived from the curated list of DR-associated mutations. These grey areas highlight the presence of DR-associated regions in candidate genes. Vertical black lines indicate the frequency of mutations observed in phenotypically resistance isolates. Internal black lines show co-occurring mutations both within and between genes. The thickness of these lines is proportional to the frequency of the mutations appearing together.

(F) Loci involved in ethionamide resistance and mutations observed in ethionamide isolates



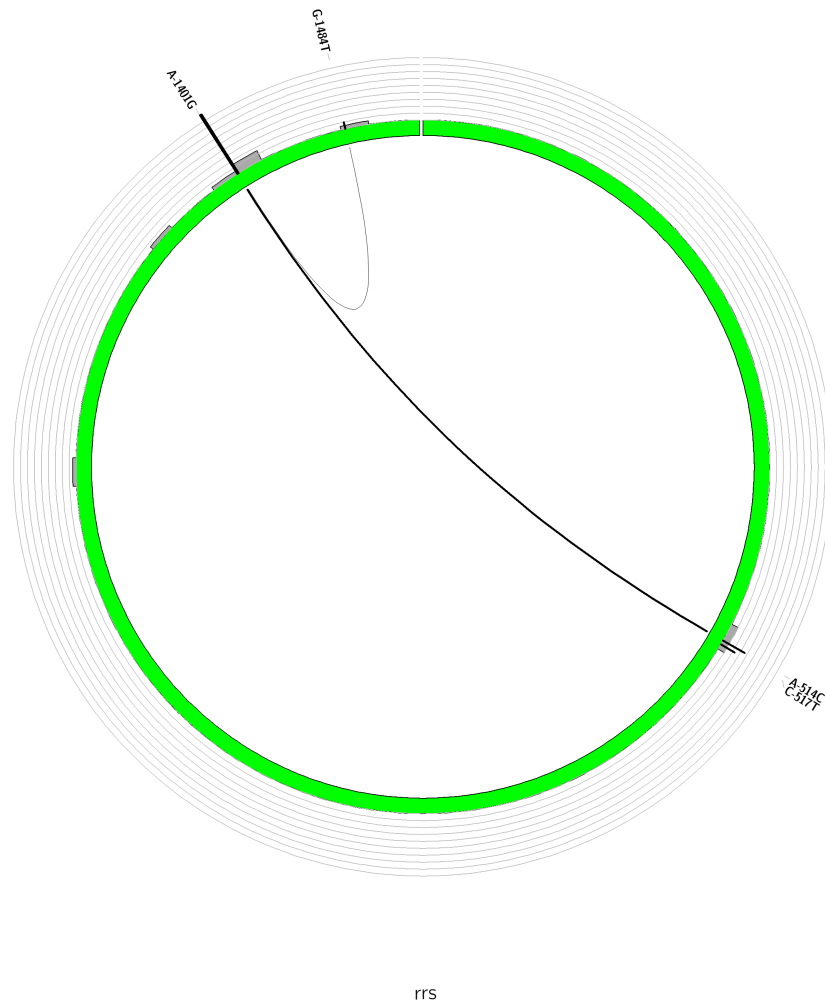
Summary of DR-associated genes and mutations in the curated library for ETH (43 variable sites, 41 SNPs and 5 indels, in 3 genes and 1 promoter) and mutations observed in phenotypically ETH resistant samples. Colour-coded bars in the Circos plot represent genes described to be involved in DR. On top of each of these bars a grey histogram shows the mutation density derived from the curated list of DR-associated mutations. These grey areas highlight the presence of DR-associated regions in candidate genes. Vertical black lines indicate the frequency of mutations observed in phenotypically resistance isolates. Internal black lines show co-occurring mutations both within and between genes. The thickness of these lines is proportional to the frequency of the mutations appearing together.

(G) Loci involved in fluoroquinolones resistance and mutations observed in ofloxacin and moxifloxacin isolates



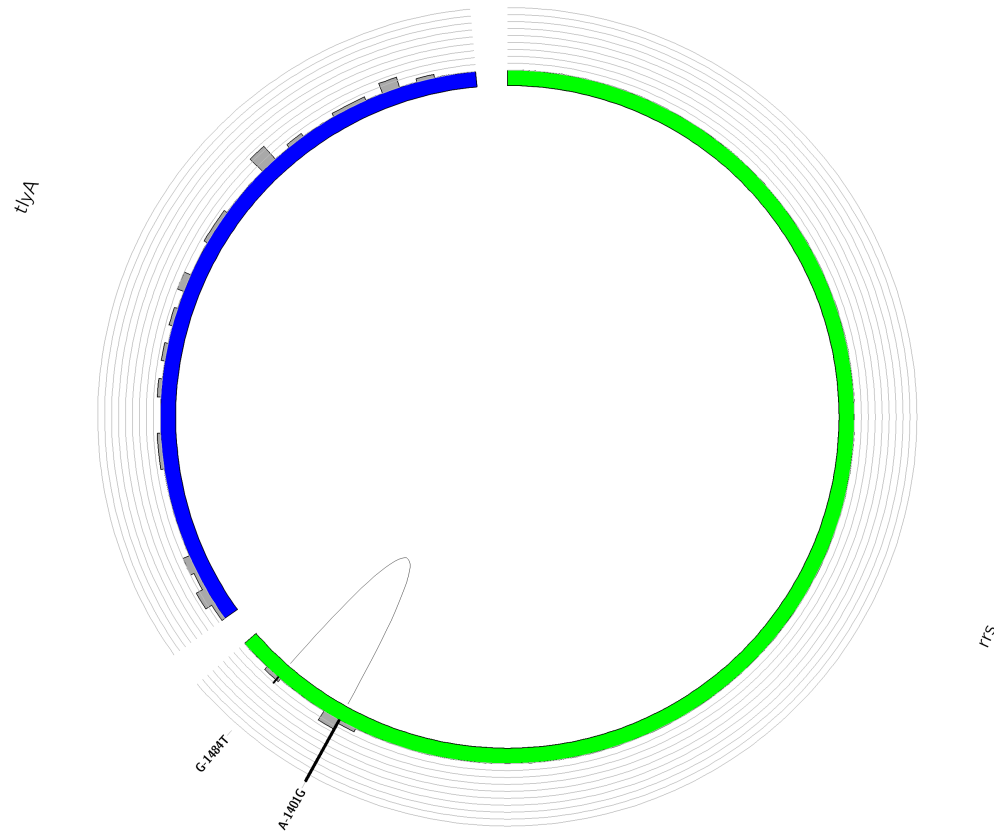
Summary of DR-associated genes and mutations in the curated library for FLQ (38 variable sites, 52 SNPs, 2 genes) and mutations observed in phenotypically FLQ resistant samples. Colour-coded bars in the Circos plot represent genes described to be involved in DR. On top of each of these bars a grey histogram shows the mutation density derived from the curated list of DR-associated mutations. These grey areas highlight the presence of DR-associated regions in candidate genes. Vertical black lines indicate the frequency of mutations observed in phenotypically resistance isolates. Internal black lines show co-occurring mutations both within and between genes. The thickness of these lines is proportional to the frequency of the mutations appearing together.

(H) Loci involved in amikacin resistance and mutations observed in amikacin isolates



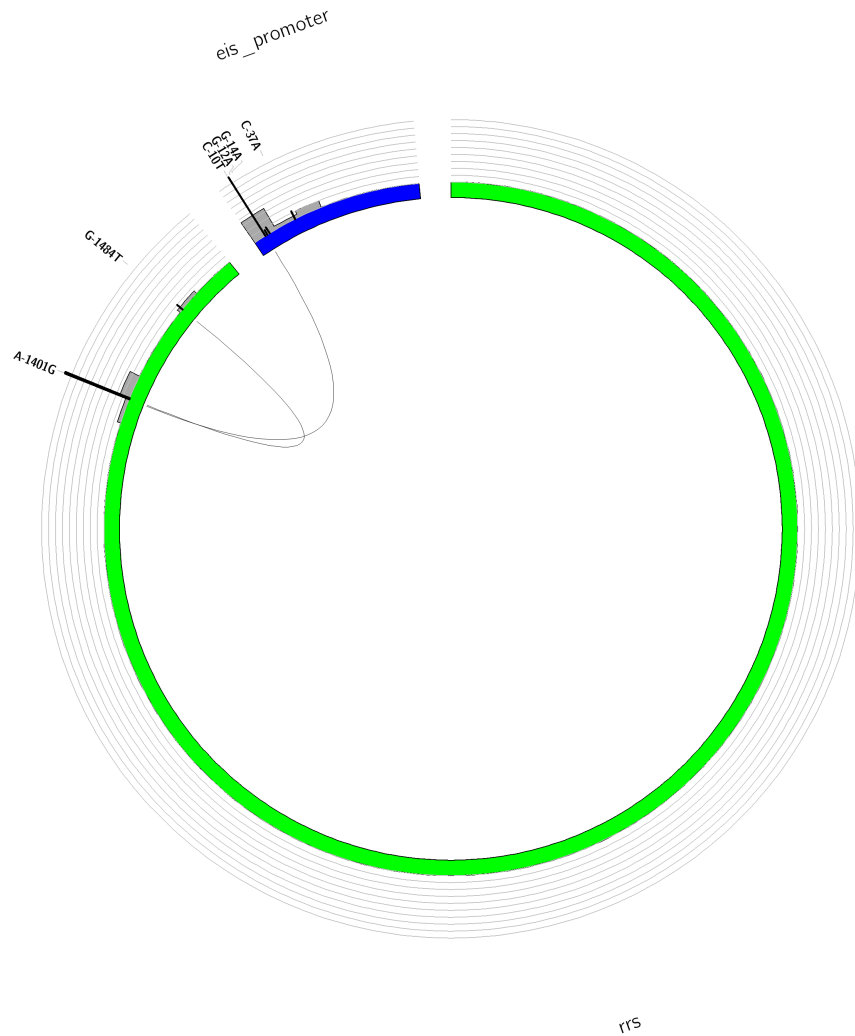
Summary of DR-associated genes and mutations in the curated library for AMK (8 variable sites, 9 SNPs, in 1 gene) and mutations observed in phenotypically AMK resistant samples. Colour-coded bars in the Circos plot represent genes described to be involved in DR. On top of each of these bars a grey histogram shows the mutation density derived from the curated list of DR-associated mutations. These grey areas highlight the presence of DR-associated regions in candidate genes. Vertical black lines indicate the frequency of mutations observed in phenotypically resistance isolates. Internal black lines show co-occurring mutations both within and between genes. The thickness of these lines is proportional to the frequency of the mutations appearing together.

(I) Loci involved in capreomycin resistance and mutations observed in capreomycin resistant isolates



Summary of DR-associated genes and mutations in the curated library for CAP (29 variable sites, 22 SNPs and 10 indels, in 2 genes) and mutations observed in phenotypically CAP resistant samples. Colour-coded bars in the Circos plot represent genes described to be involved in DR. On top of each of these bars a grey histogram shows the mutation density derived from the curated list of DR-associated mutations. These grey areas highlight the presence of DR-associated regions in candidate genes. Vertical black lines indicate the frequency of mutations observed in phenotypically resistance isolates. Internal black lines show co-occurring mutations both within and between genes. The thickness of these lines is proportional to the frequency of the mutations appearing together.

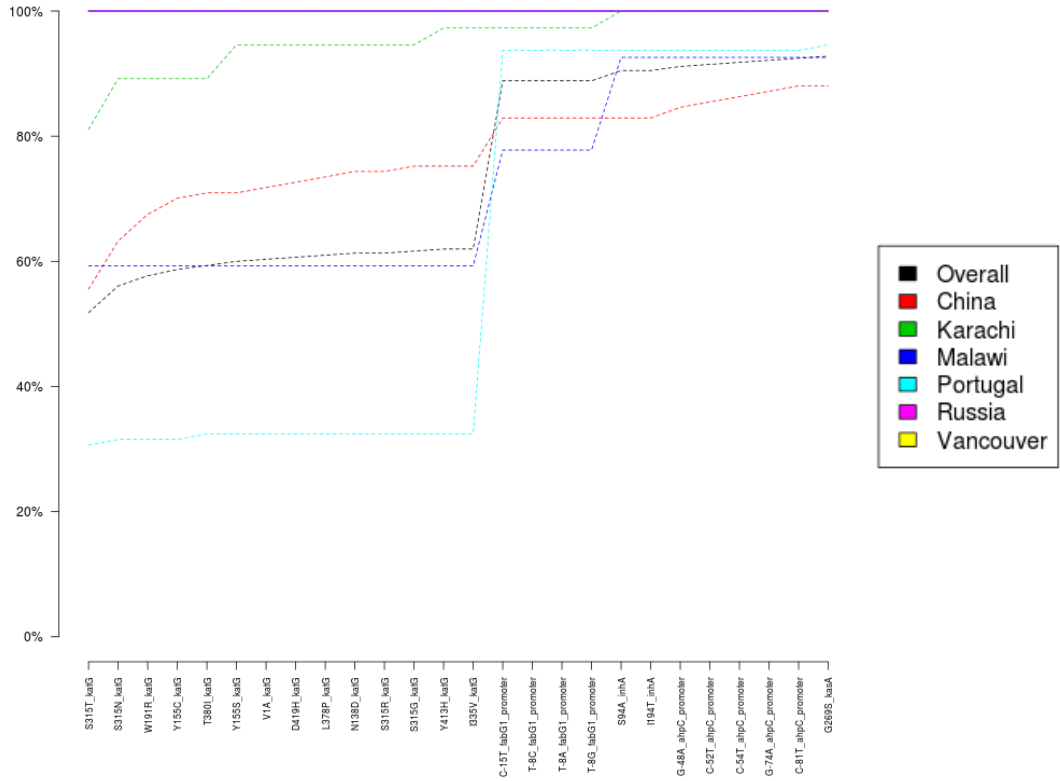
(J) Loci involved in kanamycin resistance and mutations observed in kanamycin resistant isolates



Summary of DR-associated genes and mutations in the curated library for KAN (12 variable sites, 14 SNPs, in 1 gene and 1 promoter) and mutations observed in phenotypically KAN resistant samples. Colour-coded bars in the Circos plot represent genes described to be involved in DR. On top of each of these bars a grey histogram shows the mutation density derived from the curated list of DR-associated mutations. These grey areas highlight the presence of DR-associated regions in candidate genes. Vertical black lines indicate the frequency of mutations observed in phenotypically resistance isolates. Internal black lines show co-occurring mutations both within and between genes. The thickness of these lines is proportional to the frequency of the mutations appearing together.

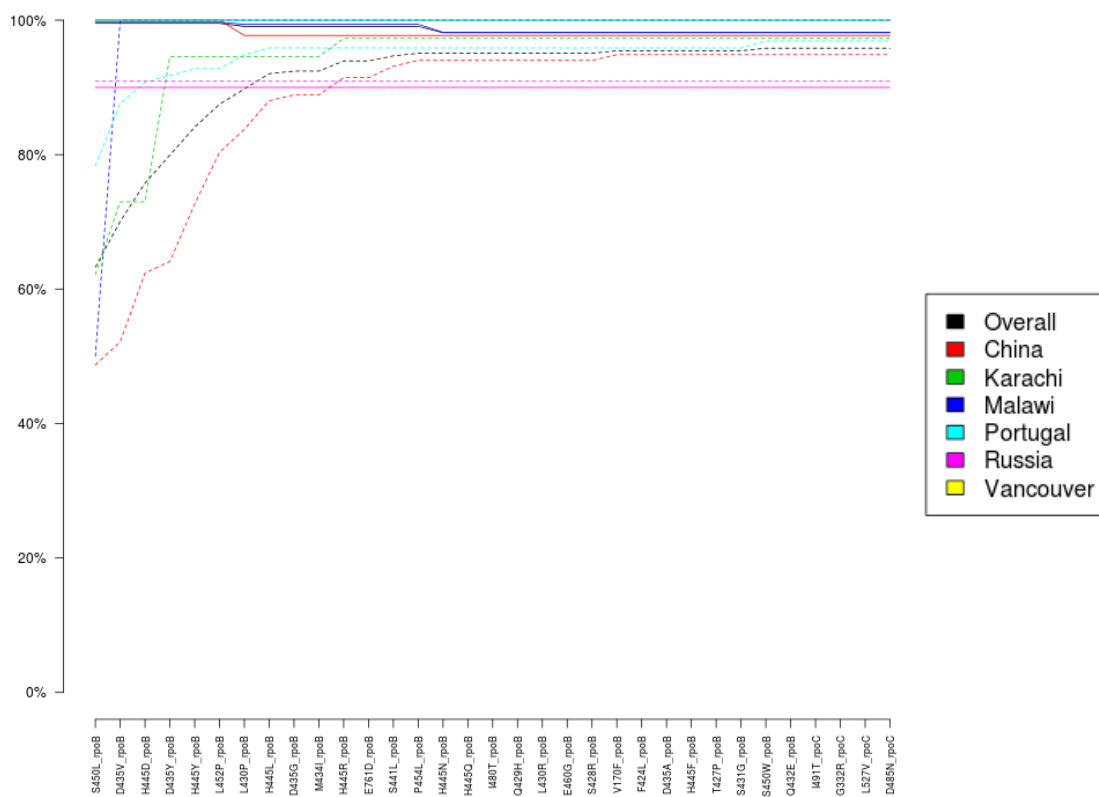
Supplementary Figure 9 Cumulative sensitivity and specificity of drug resistance markers

(A) Cumulative sensitivity and specificity of isoniazid resistance markers



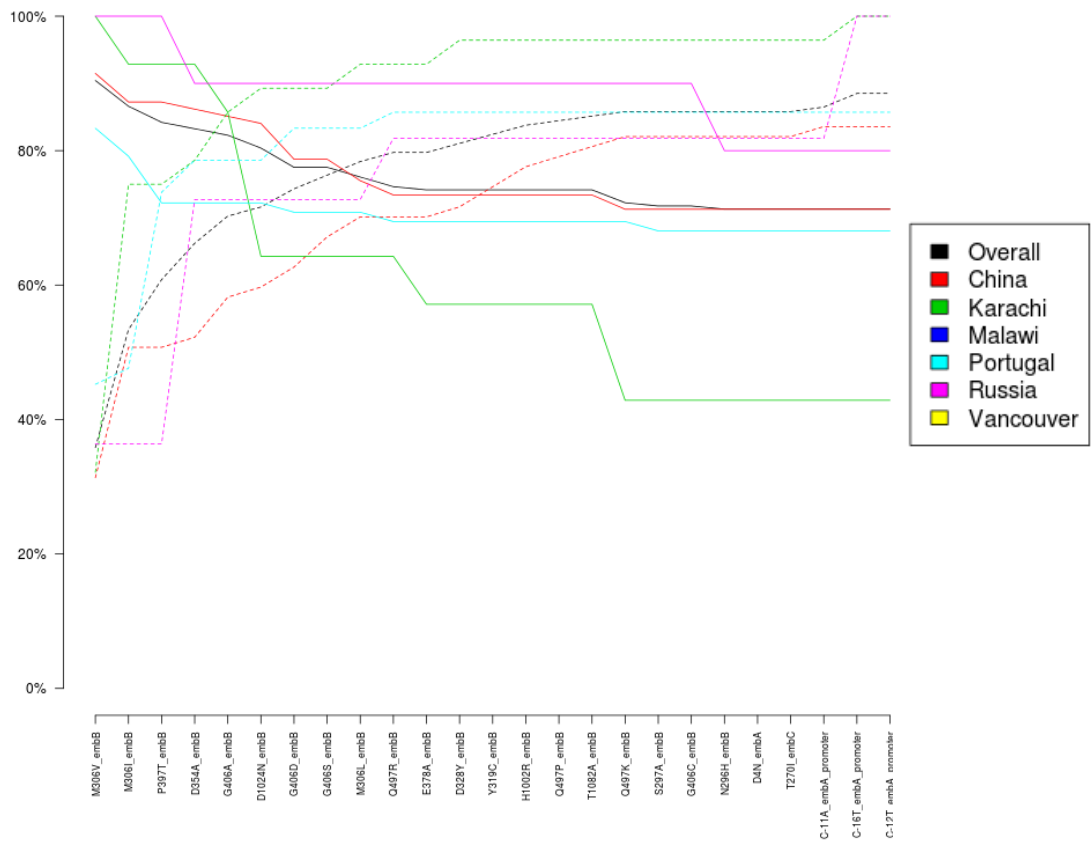
The x-axis contains DR associated mutations in the curated list observed in the overall population (WGS data set 3). Dotted lines represent sensitivity while solid lines specificity (y-axis). Lines are colour-coded by population. The plot shows the cumulative effect on sensitivity and specificity of adding a new DR mutation at a time. Mutations are ordered in the x-axis by locus and sensitivity (meaning that mutations observed more frequently are placed before in the axis).

(B) Cumulative sensitivity and specificity of rifampicin resistance markers



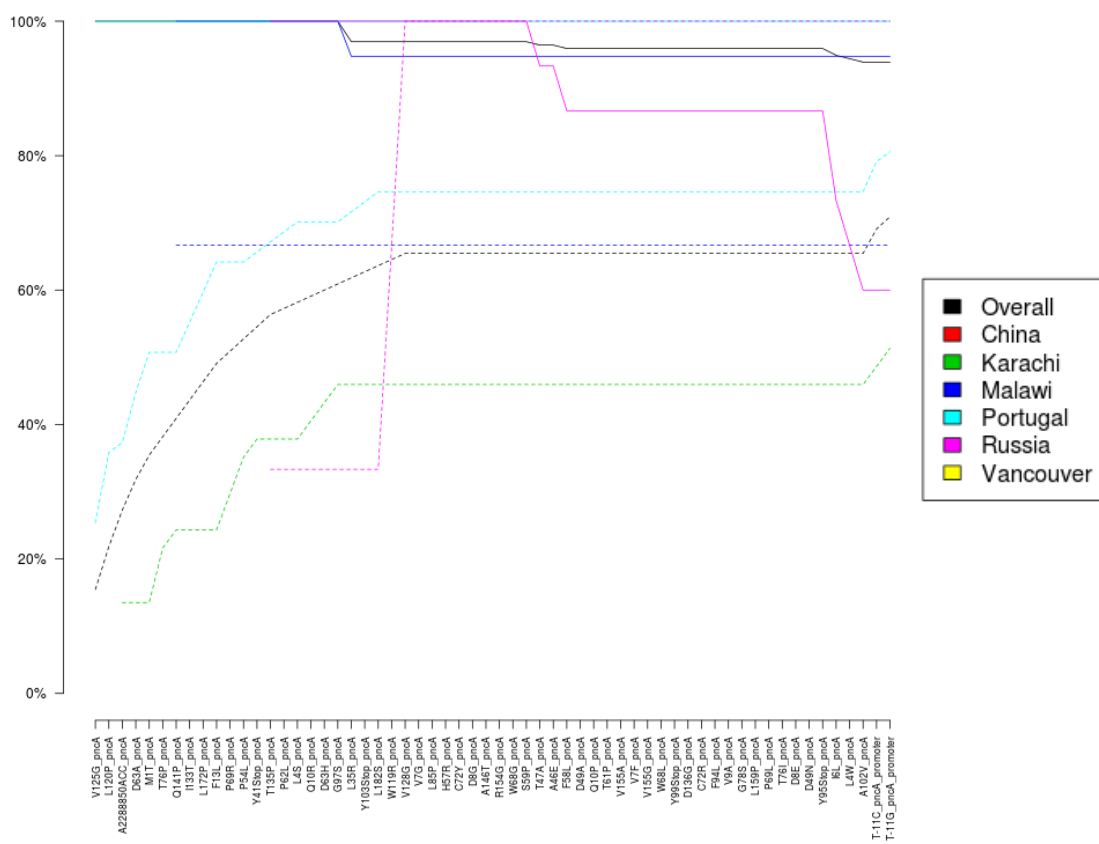
See footnote in Supplementary Figure 9A for a description of this plot.

(C) Cumulative sensitivity and specificity of ethambutol resistance markers



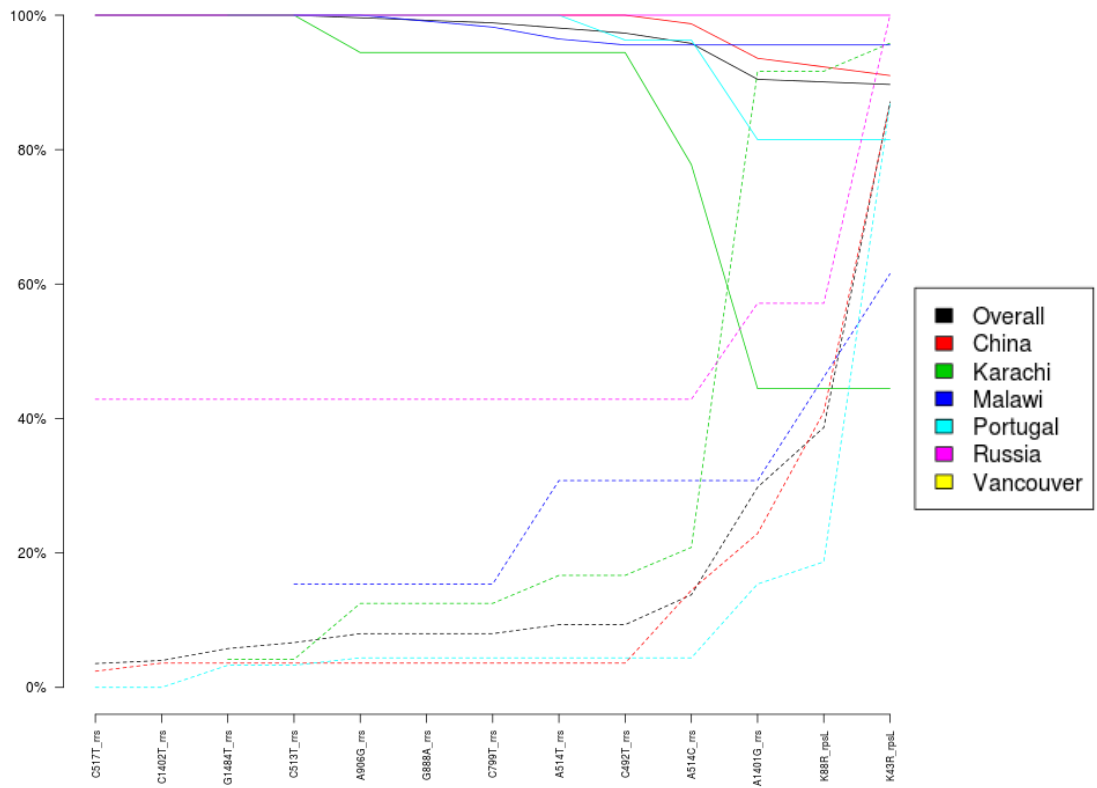
See footnote in Supplementary Figure 9A for a description of this plot.

(D) Cumulative sensitivity and specificity of pyrazinamide resistance markers



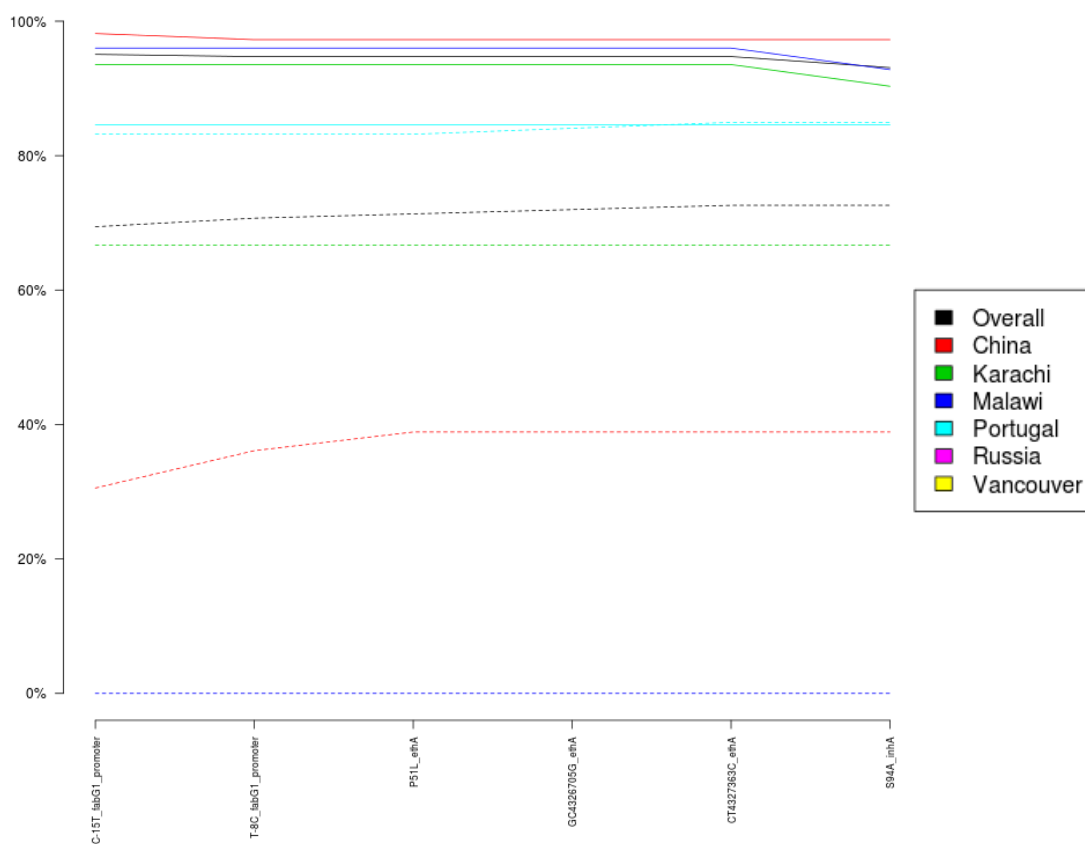
See footnote in Supplementary Figure 9A for a description of this plot.

(E) Cumulative sensitivity and specificity of streptomycin resistance markers



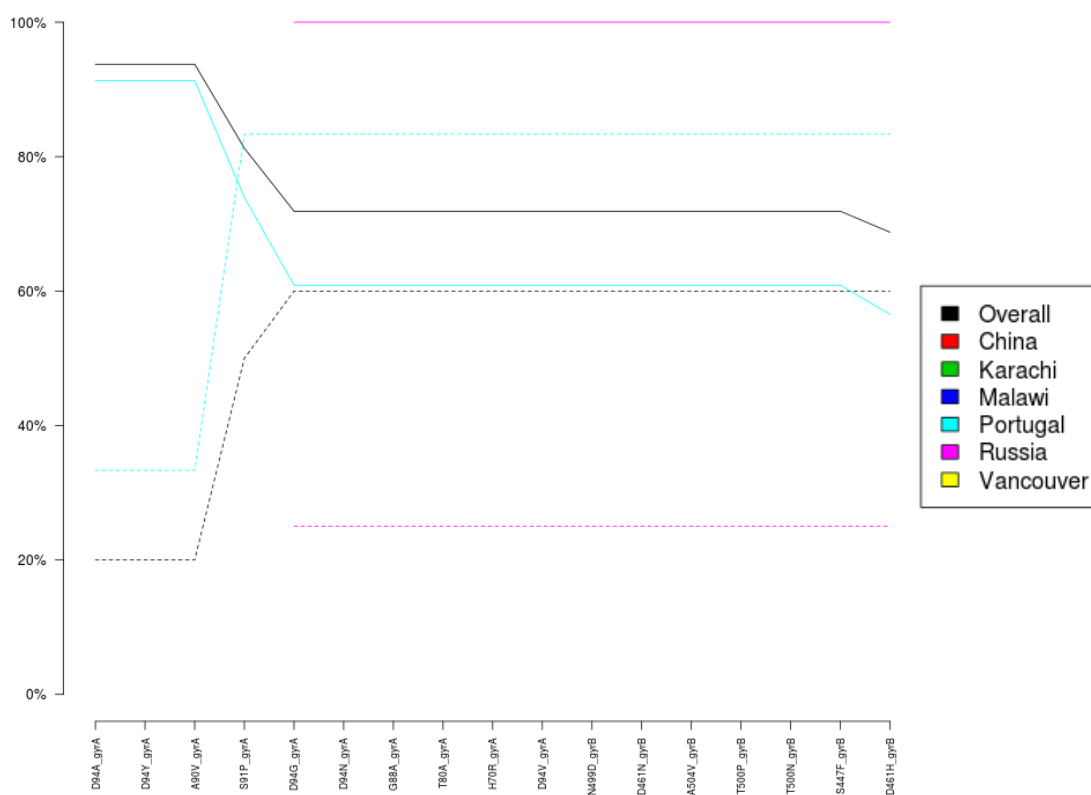
See footnote in Supplementary Figure 9A for a description of this plot.

(F) Cumulative sensitivity and specificity of ethionamide resistance markers



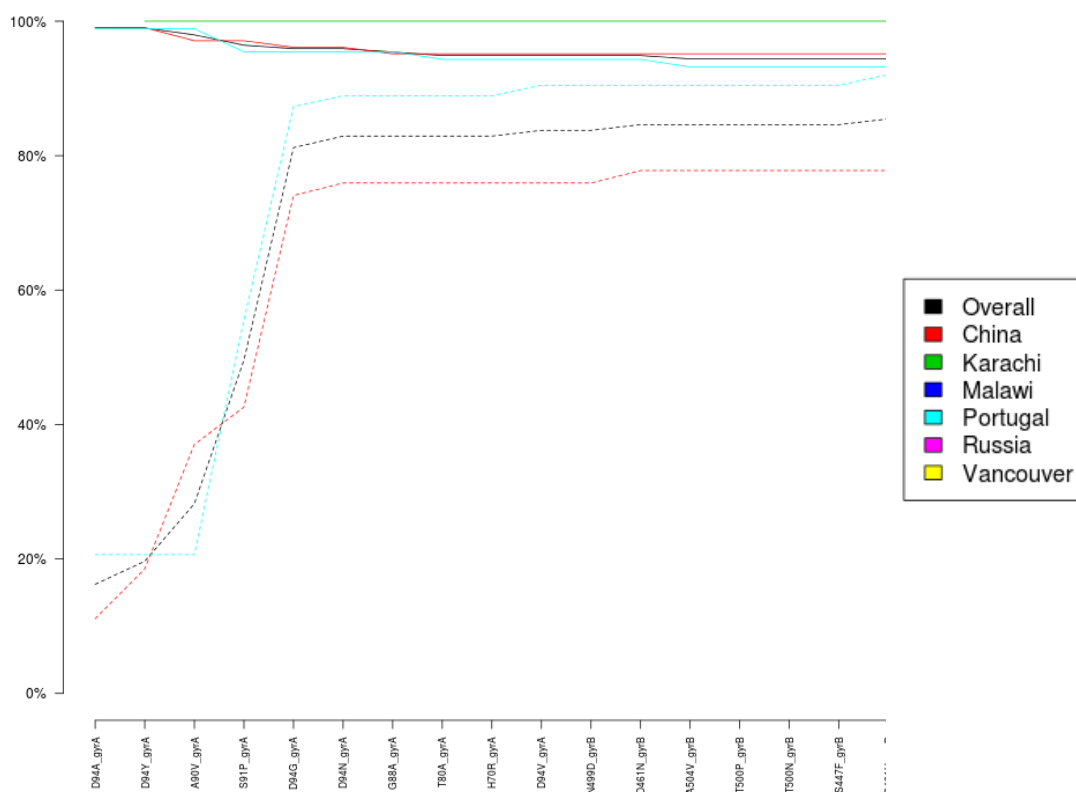
See footnote in Supplementary Figure 9A for a description of this plot.

(G) Cumulative sensitivity and specificity of moxifloxacin resistance markers



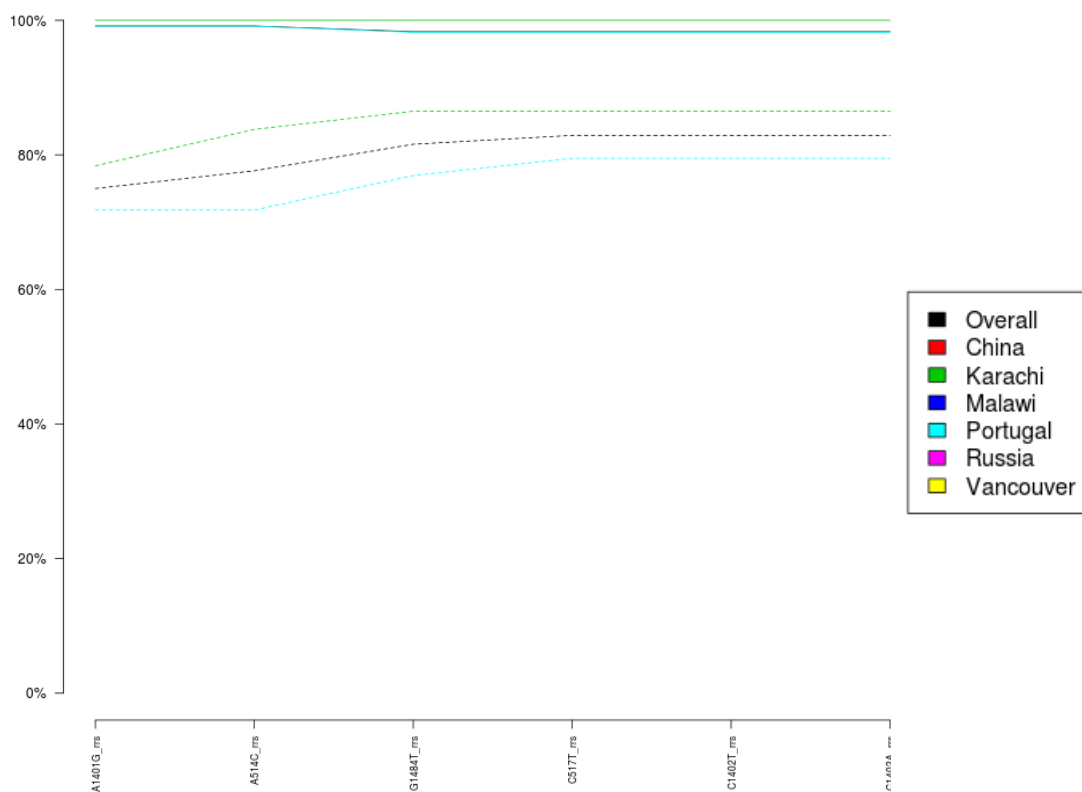
See footnote in Supplementary Figure 9A for a description of this plot.

(H) Cumulative sensitivity and specificity of ofloxacin resistance markers



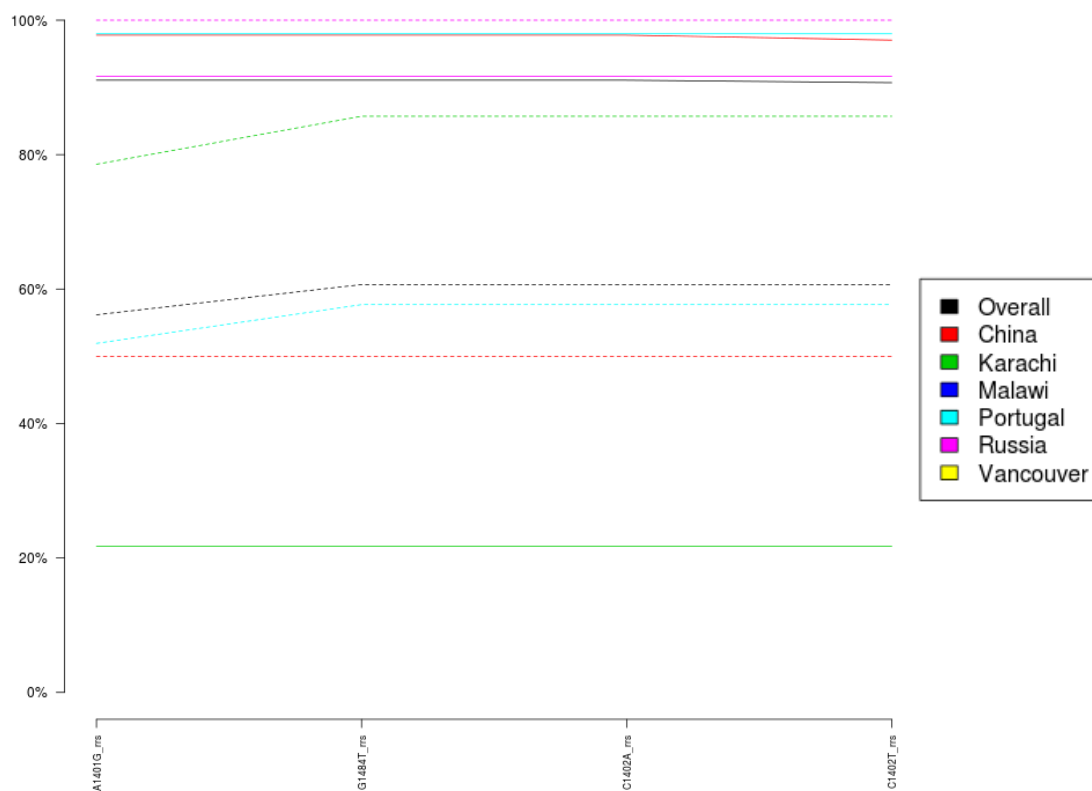
See footnote in Supplementary Figure 9A for a description of this plot.

(I) Cumulative sensitivity and specificity of amikacin resistance markers



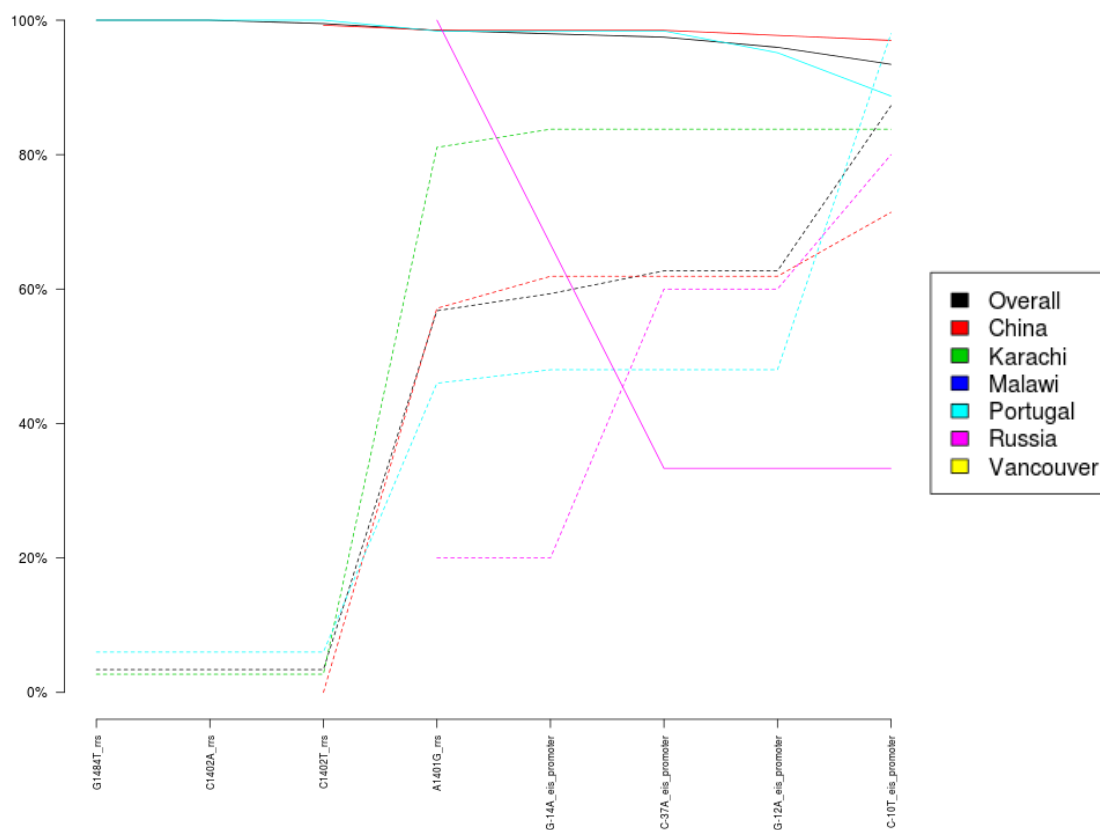
See footnote in Supplementary Figure 9A for a description of this plot.

(J) Cumulative sensitivity and specificity of capreomycin resistance markers



See footnote in Supplementary Figure 9A for a description of this plot.

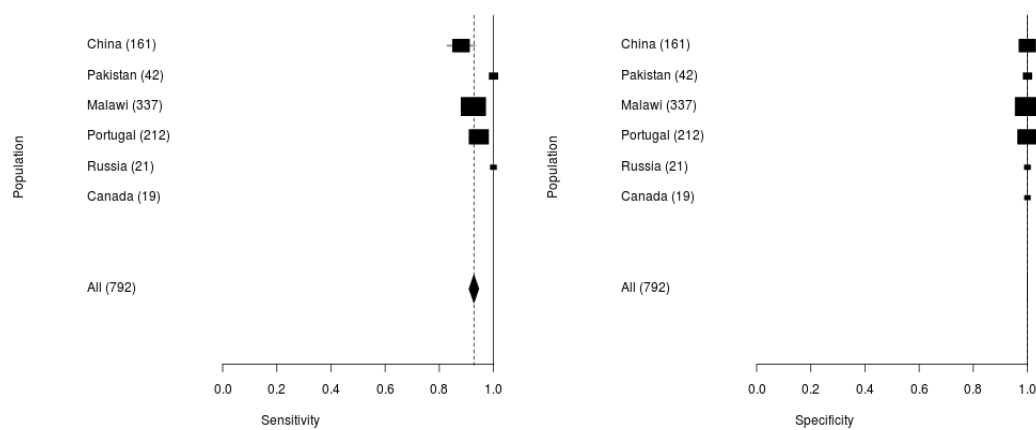
(K) Cumulative sensitivity and specificity of kanamycin resistance markers



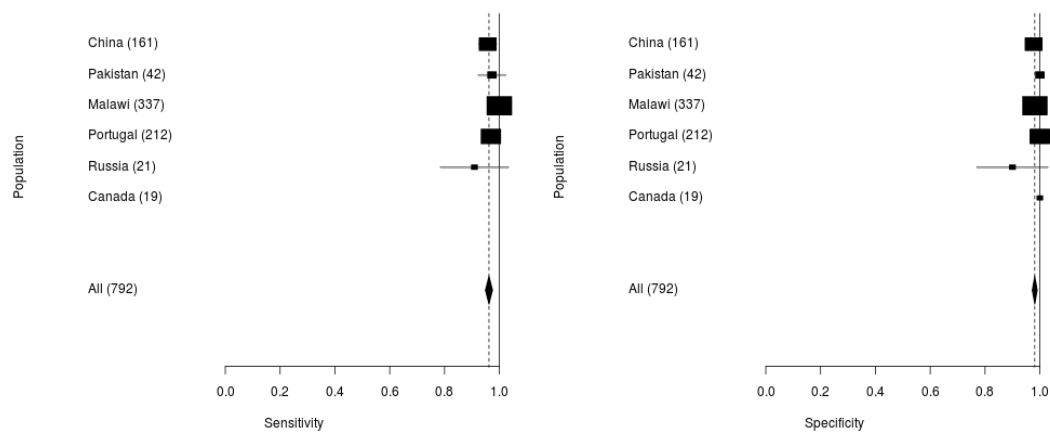
See footnote in Supplementary Figure 9A for a description of this plot.

Supplementary Figure 10 Diagnostic accuracy across populations

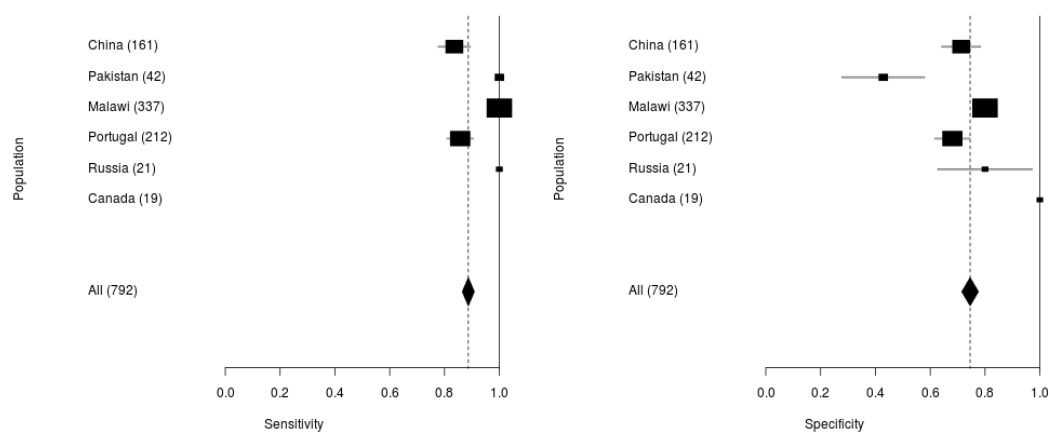
a) Isoniazid



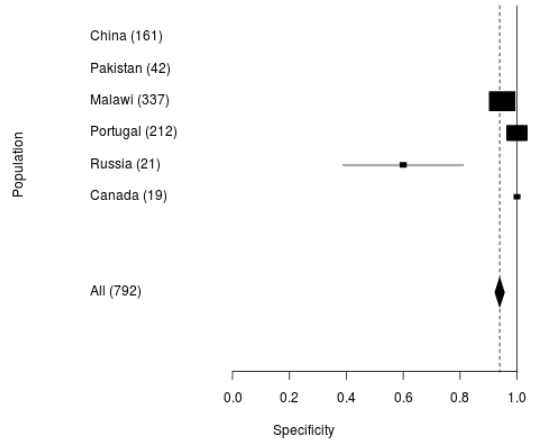
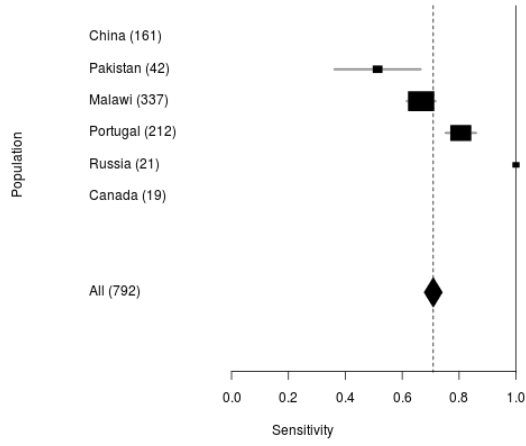
b) Rifampicin



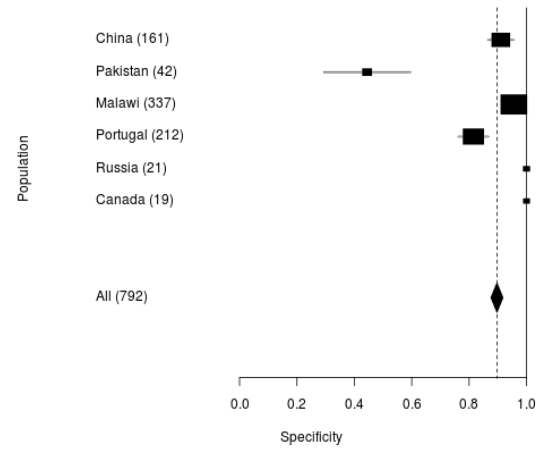
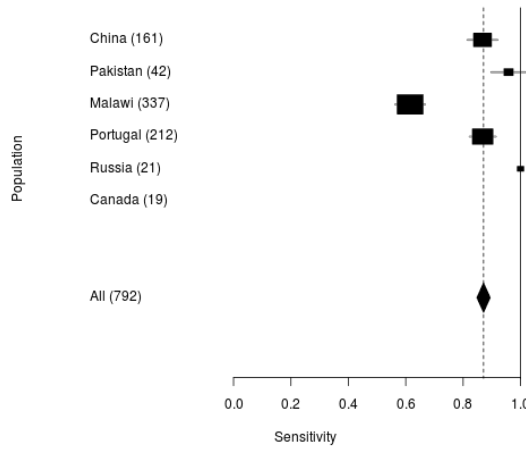
c) Ethambutol



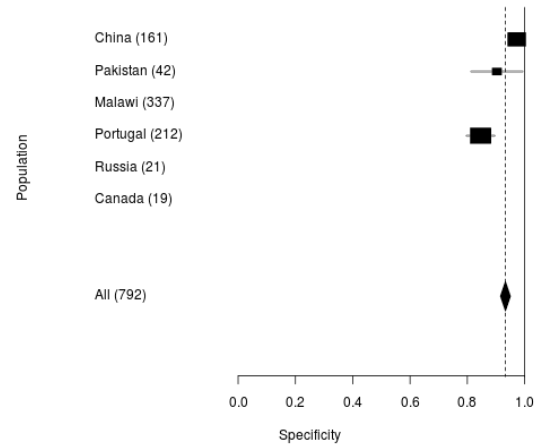
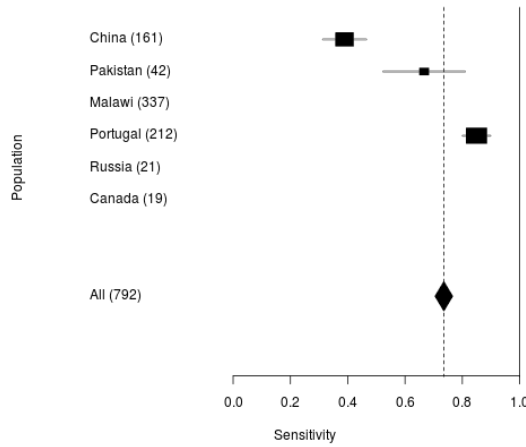
d) Pyrazinamide



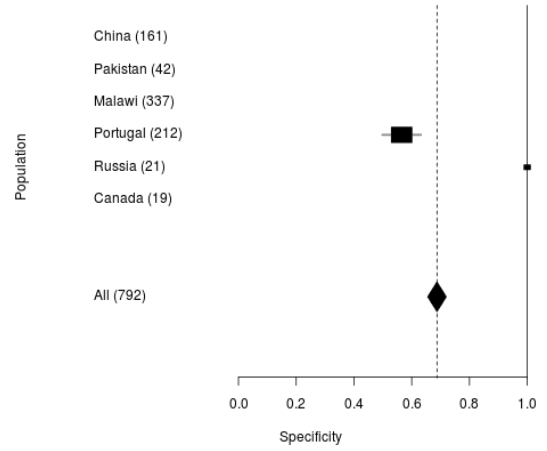
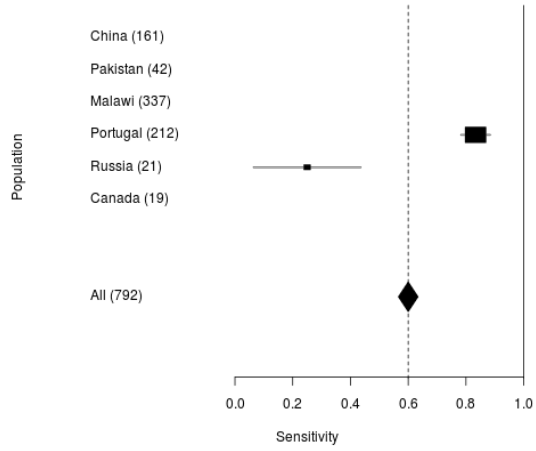
e) Streptomycin



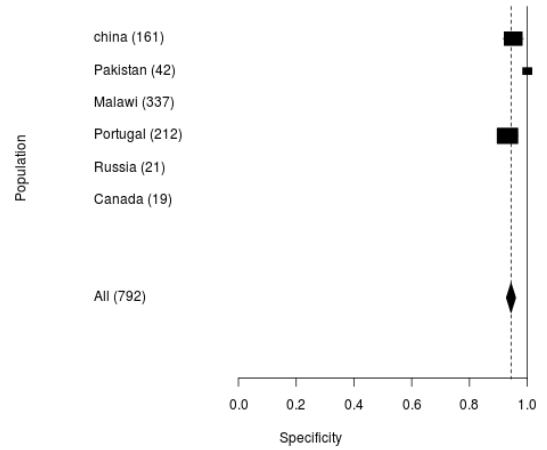
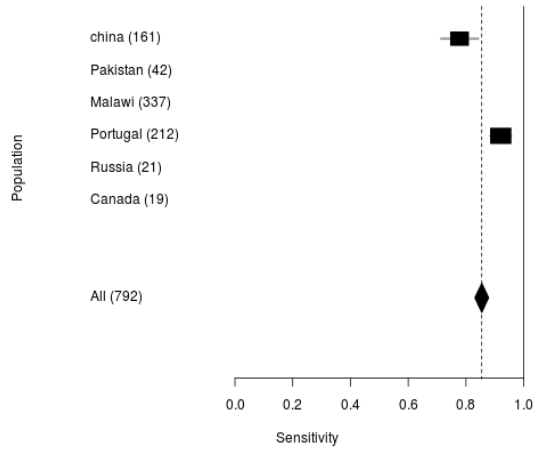
f) Ethionamide



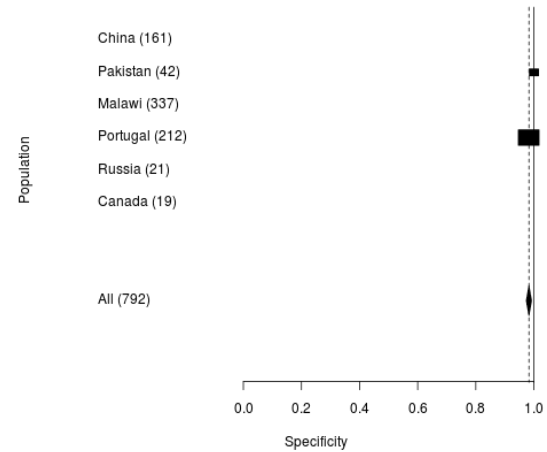
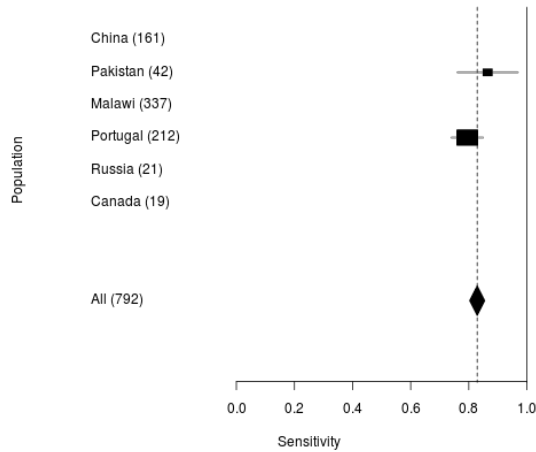
g) Moxifloxacin



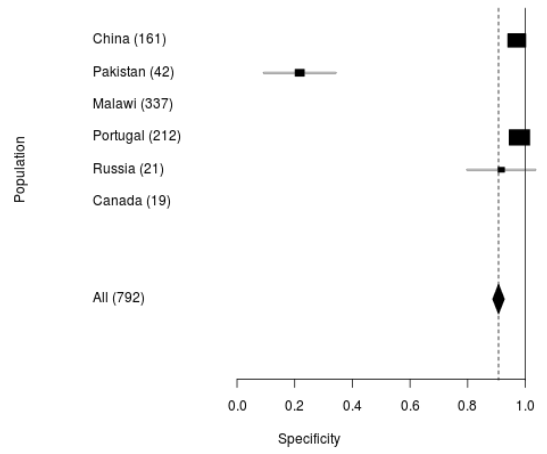
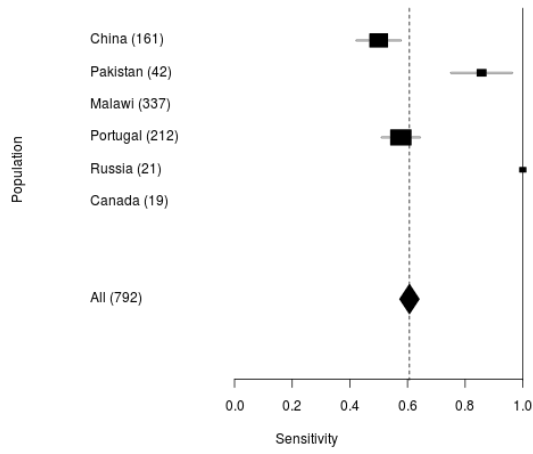
h) Ofloxacin



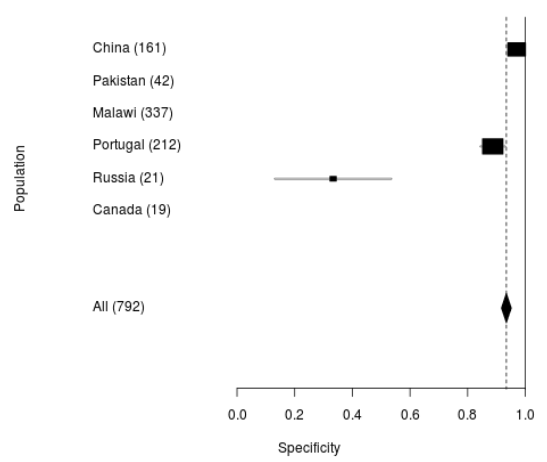
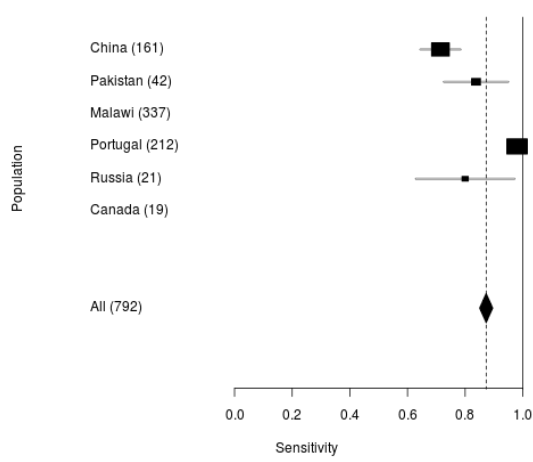
i) Amikacin



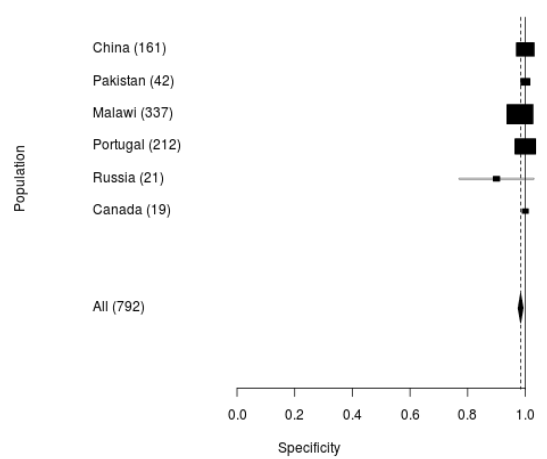
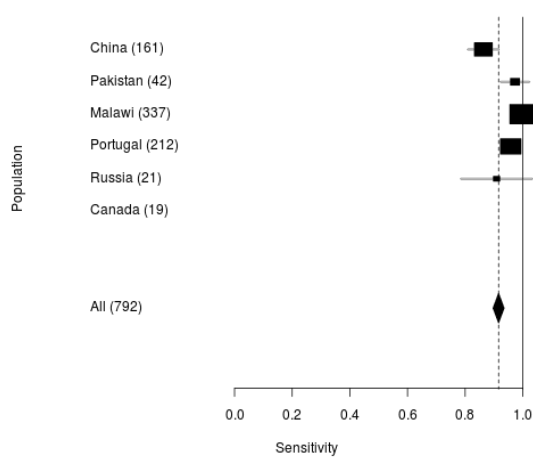
j) Capreomycin



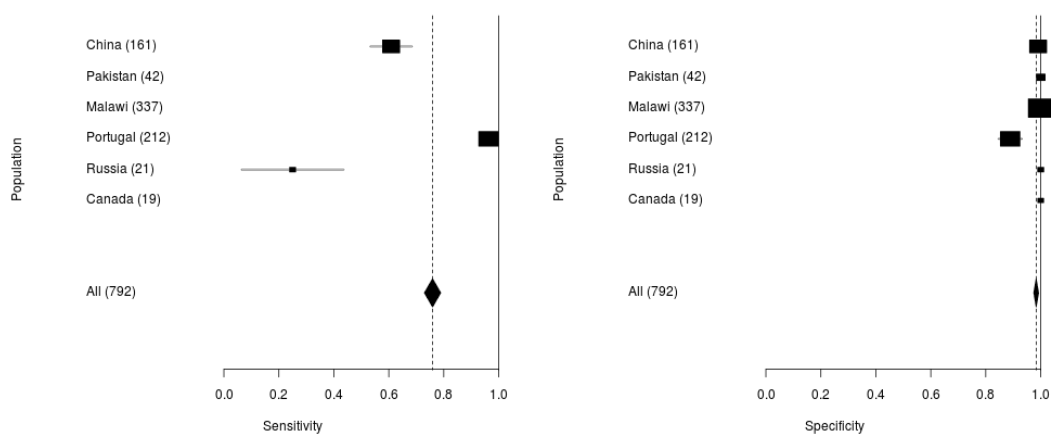
k) Kanamycin



l) MDR



m) XDR



The sensitivity and specificity (i.e. diagnostic accuracy) of DR mutations in the curated list is calculated for each drug, both within each population and overall. The point estimates are represented by solid rectangles with size proportional to the population size, where horizontal lines represent the 95% confidence intervals. The overall estimate is represented by a diamond with width representing the 95% confidence interval. Dotted vertical lines are drawn at the overall estimates. The data presented in this supplementary figure correspond to that of Table 4.2.